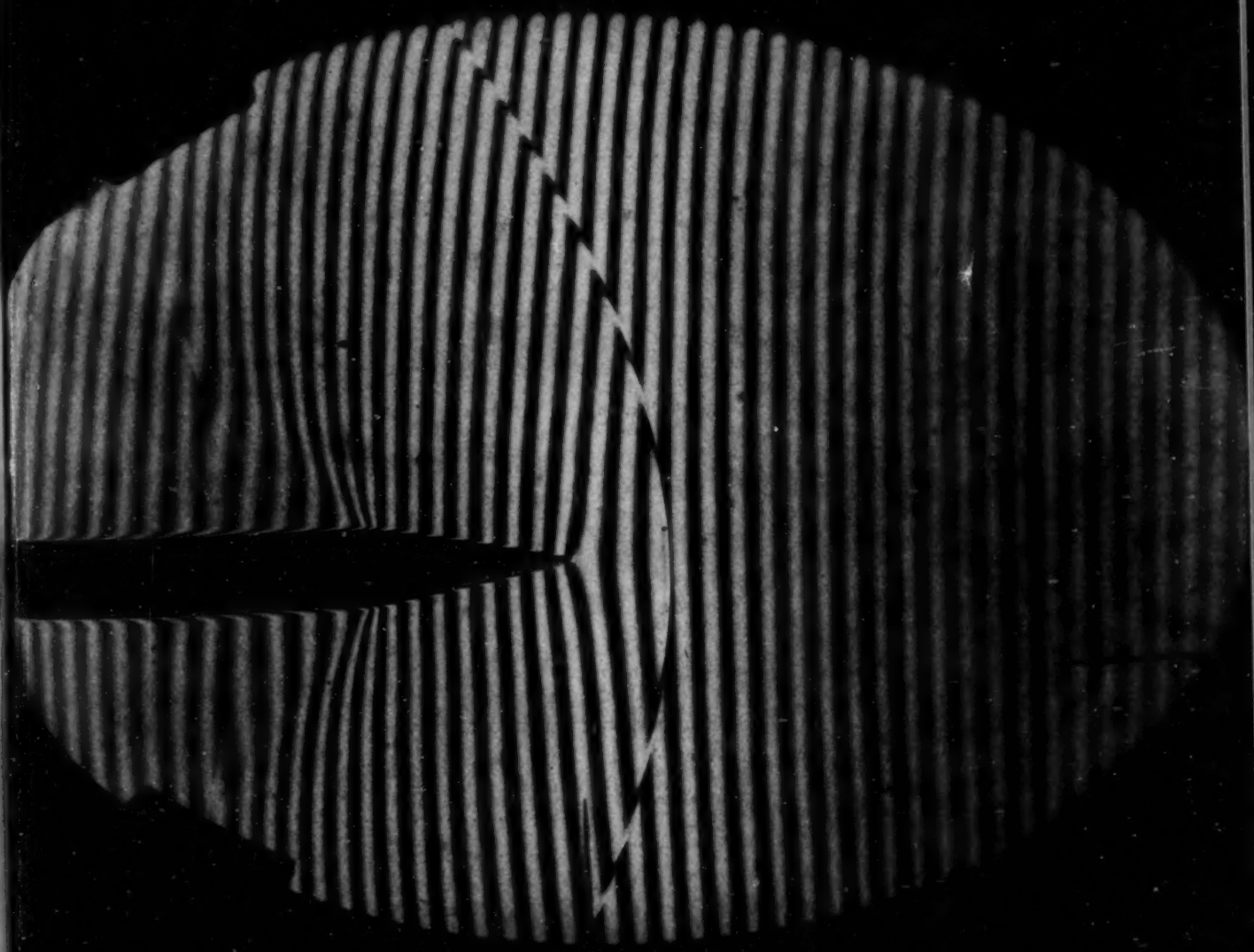
JOURNAL OF APPLIED PHYSICS



JOURNAL OF APPLIED PHYSICS

Vol. 20, No. 4 APRIL, 1949

In This Issue

	Page
Investigations of a Bearing for Small Angular Deflections without Any Friction and with Negligible Restoring Torque — Helmut Schlitt	297
On Supersonic Flow Patterns — M. M. Munk	302
Growth and Structure of Thin Metallic Films — Henry Levinstein	306
Use of a Mechanical Harmonic Synthesizer in Electrical Network Analysis — S. Leroy Brown and Chester M. McKinney	316
Theory of Hardness and Measurements Applicable to Contact Problems — Else Holm, Ragnar Holm, and Erle I. Shobert II	319
Radiation Resistances of Loaded Antennas — Richard C. Raymond and Wayne Webb	328
Current Distributions on Some Simple Antennas — Wayne Webb and Richard C. Raymond	330
On Theoretical Signal-to-Noise Ratios in F-M Receivers: A Comparison with Amplitude Modulation — David Middleton	334
Effect of Surface Roughness on Eddy Current Losses at Microwave Frequencies — Samuel P. Morgan, Jr.	352
Flash Radiography Applied to Ordnance Problems — J. C. Clark	363
Point-to-Plane Impulse Corona — Donald Baker Moore and William N. English	370
Thermoelectric Experiments with Extreme-Pressure Lubricants — Robert Schnurmann	376
Modification of Silica Replica Technique for Study of Biological Membranes and Application of Rotary Condensation in Electron Microscopy — F. Heinmets	384
Enumeration of Fourier Phase Permutations Using Electronic Analog Machines — A. D. Booth	388
On Surface-Fitting in Three Variables — Morris Morduchow	390
Reflection and Transmission of Electromagnetic Waves by a Spherical Shell — Herbert B. Keller and Joseph B. Keller	393
Microwave Antennas and Dielectric Surfaces — R. M. Redheffer	397
Letters to the Editor:	
Aberration Phenomenon in Electrostatic Lenses — Shirley Harrison	412
Absorption of Microwaves in Gases — George H. Millman and Richard C. Raymond	413
Discussion of "Geometrical Factors Affecting X-Ray Spectrometer Maxima" — Roy C. Spencer	413
Cathodic Vacuum Etching of Metals — Don M. McCutcheon	414
Table of Interplanar Spacings for Different Target Materials for the Back Reflection Region — Suzanne van Dijke Beatty	415
Pulse Emission Decay Phenomenon in Oxide-Coated Cathodes — G. R. Feaster	415
Running Striations and Multiple Electron Temperatures — N. L. Oleson and C. G. Found	416
Spherical Particles for Electron Microscopy — G. David Scott	417
Transonic Flow in a Shock Tube — D. K. Weimer, C. H. Fletcher, and W. Bleakney	418
Index to Advertisers	xxxviii

BOARD OF EDITORS
Elmer Hutchisson, *Editor*

Associate Editors
R. Bowling Barnes
Otto Beeck
W. James Lyons
Philip Morrison
Paul E. Sabine
Raymond J. Seeger
Frederick Seitz
Charles S. Smith
J. R. Townsend

Publication of the American Institute of Physics

OFFICERS
George R. Harrison, *Chairman*
Wallace Waterfall, *Secretary*
George B. Pegram, *Treasurer*
John T. Tate, *Adviser on Publications*

GOVERNING BOARD
Luis W. Alvarez
J. W. Beams
Wallace R. Brode
E. U. Condon
W. F. Fair, Jr.
Harvey Fletcher
George R. Harrison
W. V. Houston
Paul Kirkpatrick
Paul E. Klopsteg
F. W. Loomis
Philip M. Morse
J. R. Oppenheimer
George B. Pegram
Harold K. Schilling
John T. Tate
Wallace Waterfall
M. W. White
A. G. Worthing

Consultant on Printing
Melvin Loos

ADMINISTRATIVE STAFF
Henry A. Barton, *Director*
Theodore Vorburger, *Business Manager*
Ruth F. Bryans, *Publication Manager*
Betty Hofstadter, *Office Manager*
Eileen B. Neuberger, *Circulation Manager*

The Journal of Applied Physics, published monthly at Prince and Lemon Sts., Lancaster, Pa., is devoted to physics in its role as the science basic to other natural sciences and to the arts and industries. Previous to 1937 this publication was known as Physics. The Journal of Applied Physics publishes editorials, news of physicists, and reviews, as well as technical papers of applied physics. Articles appearing in it are indexed regularly, according to their field, by Physics Abstracts, Chemical Abstracts, Engineering Index, Inc., Industrial Arts Index, etc.

Manuscripts should be submitted to Elmer Hutchisson, Editor, Case Institute of Technology, Cleveland 6, Ohio. The authors' institutions are requested to pay a **publication charge** of $4 per page which, if honored, entitles them to 100 free reprints. Instructions will be sent with galley proofs.

Subscriptions and orders for back numbers should be addressed to Prince and Lemon Streets, Lancaster, Pa. or to the American Institute of Physics, 57 East 55 Street, New York 22, N. Y.

Subscription Price	U.S. and Canada	Elsewhere
To members and associate members of the American Institute of Physics	$7.00	$ 8.00
To all others	9.00	10.00

Back Number Prices
Complete set: Vol. 1, July, 1931–Vol. 19, 1948—$140.00
Yearly back number rate: $10.00
Single copies: $1.00 each.

Changes of address should be sent to the Business Manager, American Institute of Physics; **corrected proofs** should be sent to the Publication Manager, American Institute of Physics.

Advertising rates supplied on request. Orders, advertising copy, and cuts should be sent to the American Institute of Physics.

Entered as second class matter January 22, 1937, at the Post Office at Lancaster, Pennsylvania, under the Act of March 3, 1879.
Accepted for mailing at the special rate of postage provided for in the Act of February 28, 1925, authorized May 2, 1932.

Copyright 1949, by the American Institute of Physics.

Journal
of
Applied Physics

Volume 20, Number 4 April, 1949

Investigations of a Bearing for Small Angular Deflections without Any Friction and with Negligible Restoring Torque

HELMUT SCHLITT
Ordnance Research and Development Division Suboffice (Rocket), Fort Bliss, Texas
(Received March 31, 1948)

This paper deals with bearings supported on metal beams under mechanical stress in the axial direction. Because of a phenomenon related to the bending of long columns under axial compression the torque of the total arrangement can be made zero or even negative. Such a bearing can be designed with two decisive advantages, absolutely no friction, and practically no restoring torque as far as small angles of rotation are concerned. Two different types of beam bearings are treated, and their equations for precalculation are derived. Furthermore, the influences of temperature, external forces, and the angle of rotation on the torque of the bearing are investigated. Finally, a combination of two different beam bearings is shown to offer a possible compensation for these unfavorable effects.

THE solution of the problem of designing a bearing without any friction or restoring torque is for many reasons very important. For example, the application of a gyro in control devices requires air or liquid bearings to reduce friction and so to obtain highest accuracy. In many cases the angular deflection of such a bearing can be kept very small, but the requirement that there be practically no friction or restoring force has to be met.

The following description deals with a bearing—the basic idea of which is related to the laws of bending of long columns under axial compression. It is well known that a column, because of its length, cannot be held in a straight line if the axial load exceeds a certain limit. At this very limit of the axial compression a slight sidewise bending occurs perpendicular to that axis of the cross section about which the moment of inertia is the least. Within certain narrow limits neither the amount nor the sign of that lateral deflection is defined. This means that within those limits any deflection of the column's center part can occur with practically no additional bending force. This effect, suitably transferred to the design of a bearing, can be used to yield a rotation within a small range of angle without applying any force for turning.

Figure 1 shows the principal arrangement of such a bearing. On a circle A, with the radius R_1 attached to the movable part of the bearing, there are three metal rods rigidly secured in the radial directions. The other ends of those metal beams are fixed on the circle B belonging to the stationary part of the bearing. The distance of these ends from the center of motion O of the bearing is R_2. The three rods are somewhat displaced in the axial direction of the bearing so that they will not disturb each other. In stressing one of them by the spring S (Fig. 1) all three are under equal tensile stress.

A suitable choice of the tensile force, P_s, can produce a phenomenon similar to the above mentioned of a column under axial compression. This means that in the case of the bearing the restoring moment between the two circles A and B can become zero or even negative. Physically it can be explained as follows: By turning of circle A the fixed ends of the metal beams move on a circle with

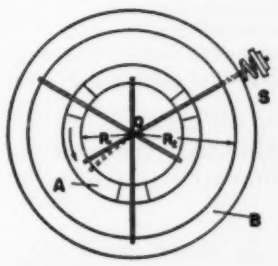


FIG. 1. Principal arrangement of the bearing.

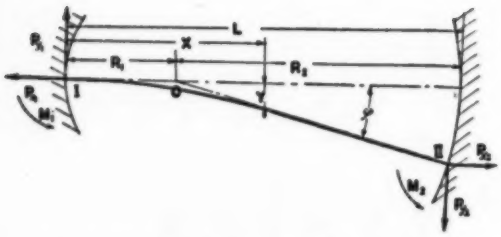


FIG. 2. Elastic curve of a single metal beam.

the radius R_1 (see dash line in Fig. 1, marking the position of one of those metal beams). It is evident that because of the tensile force of the beams a negative restoring moment acts on the circle A and increases with the angle of rotation. On the other hand, the beams give a positive restoring moment because of their natural rigidity. Therefore it seems evident that for a suitable choice of all proportions, especially of P_s, the total restoring moment of the bearing can be made zero or even negative.

The following discussion gives a short quantitative representation of the relations wherefrom all consequent requirements for the design of such bearings can be derived.

For this purpose the elastic curve of a single metal beam is represented in Fig. 2. The tangents to the elastic curve at the ends I and II correspond to both radii crossing in the center of motion O. In the case in question the circle B has been turned through the angle φ out of its original position. Also, since in the practical case of application the angle φ will always be very small (less than 1°), the curvature of the beam can be approximated for the following calculation with sufficient accuracy by the second differential quotient of the elastic curve. Then the following differential equation of the beam under axial tensile stress results

$$(d^2y/dx^2)-(P_x/I\cdot E)\cdot y = -(P_y/I\cdot E)\cdot x+(M_1/I\cdot E), \quad (1)$$

where

I = moment of inertia of the cross section of the beam with respect to the neutral axis. The following considerations are based partially on a rectangular cross section with the sides a and b ($a \gg b$). The side a is parallel to the axis of the bearing. The moment of inertia of this cross section with respect to the neutral axis parallel to a equals $a\cdot b^3/12$.
E = modulus of elasticity,
P_x = force in the direction x,
P_y = force in the direction y,
M_1 = bending moment of the beam at point I, and
M_2 = bending moment of the beam at point II.

Determining the constants of integration by the boundary conditions of the problem, the solution of the above differential equation can be written as follows:

$$y=-(P_y/P_x)[(\lambda/\cosh\alpha)\cdot\sinh((x/\lambda)+\alpha) - x-\lambda\cdot\tanh\alpha]. \quad (2)$$

The letter λ represents

$$\lambda=(E\cdot I/P_x)^{0.5}. \quad (3)$$

The angle α can be expressed by the relation

$$\tanh\alpha=\frac{R_1/\lambda-\sinh(L/\lambda)+R_2/\lambda\cdot\cosh(L/\lambda)}{\cosh(L/\lambda)-R_2/\lambda\cdot\sinh(L/\lambda)-1}. \quad (4)$$

Knowing the equation of the elastic curve, the moment at point I of the beam with respect to the center of motion O can easily be calculated. Therefore the following equation for the restoring torque results:

$$M_0=\frac{\cosh\alpha\cdot(R_1+\lambda\cdot\tanh\alpha)}{\cosh(L/\lambda+\alpha)-\cosh\alpha}\cdot P_x\cdot\varphi. \quad (5)$$

That moment becomes zero under the condition

$$\tanh\alpha=-R_1/\lambda. \quad (6)$$

Upon introduction of this value for $\tanh\alpha$ into Eq. (4), the condition for disappearing torque M_0 reads

$$\tanh(L/\lambda)=(L/\lambda/(R_1\cdot R_2/L^2)\cdot(L/\lambda)^2+1). \quad (7)$$

The solution of this equation for values of $L/R_1\geqq 3$ can be represented practically with sufficient accuracy by

$$\lambda=R_1. \quad (8)$$

Using Eq. (3), this can be written as follows:

$$R_1^2=E\cdot I/P_x. \quad (9)$$

For small angles φ the force component P_x can be approximated with sufficient accuracy by the total force P_s. Introducing the unit stress s instead of P_s, and referring to the above rectangular cross section of the beam with the sides a and b ($a \gg b$),

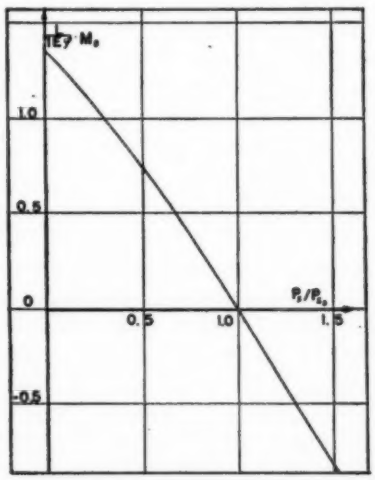


FIG. 3. Torque of a single beam as a function of the tensile force.

Eq. (9) can be written

$$R_1^2=(E\cdot b^2/12\cdot S). \tag{10}$$

Equation (10) represents the actual equation for calculation of such a beam bearing. To be safe from the effect of shocks the value of s must be chosen low with respect to the tensile strength.

Furthermore, it is worth while to see how the total moment as a function of P_x (approximately P_s) changes. To answer that question Eqs. (4) and (5) were evaluated for the special case of design $L/\lambda=3.15$. This value, as shown below, is chosen because of independence of temperature of a set of two different beam bearings. Figure 3 shows the torque curve plotted *versus* P_s/P_{s0} and calculated according to Eq. (5). P_{s0} signifies the tensile force P_s when $M_0=0$—that means in the case of exact torque compensation.

Moreover, the slope of this curve at the point $M_0=0$ can be calculated by Eq. (5). Under the assumption that $\tanh\alpha\approx 1$ it results in

$$-(dM_0/dP_{s(P_s=P_{s0})})=(\lambda/2)(\varphi/(1-\exp(-L/\lambda)))=0.15(E/S)^{0.5}\cdot b\cdot\varphi. \tag{11}$$

Equation (11) provides for precalculation of the torque due to a small deviation of the tensile force P_s from its prescribed value P_{s0}. Practically, these changes can never be avoided so that such a bearing always has a certain small torque. Especially, any changes in temperature can affect such a bearing. Moreover the tensile force P_s decreases with increasing angle φ. This is due to the fact, that the necessary beam length becomes smaller when the point II moves upon the circle with R_2 around the center O (see Fig. 2). This change in length of the elastic curve between the points I and II results in a decrease of P_s, the consequence of which, according to Fig. 3, is a positive restoring torque. Obviously such a bearing has only a negligible restoring torque in the immediate surroundings of $\varphi=0$, which, however, increases with φ to positive values as P_s varies.

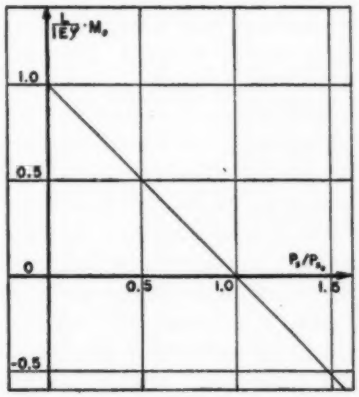


FIG. 4. Torque of a single beam as a function of the compression force.

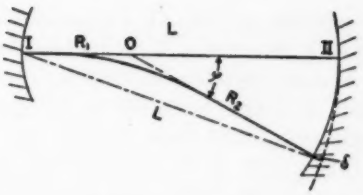


FIG. 5. Representation of the elastic curve showing the contraction of the beams with increasing torsion angle.

These disadvantages can be nearly compensated for if a second bearing under compression stress is used. The principal arrangement of such a bearing exactly equals the one shown in Fig. 1. The only difference is that the spring S supplies a compression force. The differential equation of the elastic curve of such a single beam is

$$(d^2y/dx^2)+(P_x/IE)\cdot y=-(P_y/IE)\cdot x+(M_1/IE). \tag{12}$$

With regard to the boundary conditions, the solution can be written as follows:

$$y=(P_y/P_x)[(\lambda/\cos\alpha)\cdot\sin((x/\lambda)+\alpha)-x-\lambda\cdot\tan\alpha], \tag{13}$$

where λ represents again

$$\lambda=(EI/P_x)^{0.5}. \tag{14}$$

The angle α is defined by the relation

$$\tan\alpha=\frac{R_1/\lambda+\sin(L/\lambda)+R_2/\lambda\cdot\cos(L/\lambda)}{1-\cos(L/\lambda)+R_2/\lambda\cdot\sin(L/\lambda)}. \tag{15}$$

The total torque of point I with respect to the center of motion O can be expressed by

$$M_0=\frac{\cos\alpha\cdot(R_1-\lambda\cdot\tan\alpha)}{\cos(L/\lambda+\alpha)-\cos\alpha}\cdot P_x\cdot\varphi. \tag{16}$$

This torque becomes zero under the following condition

$$R_1=\lambda\cdot\tan\alpha \quad \text{if} \quad \cos(L/\lambda+\alpha)-\cos\alpha\neq 0. \tag{17}$$

Using Eq. (15), this condition for disappearing torque can be written

$$\tan(L/\lambda)=(L/\lambda/(R_1\cdot R_2/L^2)\cdot(L/\lambda)^2-1). \tag{18}$$

Because of the periodicity of $\tan L/\lambda$ this equation has an infinite number of solutions for L/λ. They correspond to the different possibilities with respect to the shape of the elastic curve of the beam under compression stress. For the case in question only the smallest value of all the solutions for L/λ is taken into consideration. It corresponds to the condition $M_0=0$ which takes place first in case the compression force P_s increases continuously from

zero. The solution of Eq. (18) requires an assumption concerning the ratio R_2/R_1. Because the beam length of a compression bearing in most cases is rather small, it seems suitable to choose the ratio of both radii as 1. That means both radii are equal. Equation (18) yields in that case for L/λ the value 1.72. The torque Eq. (16) yields the curve shown in Fig. 4 as a function of P_s/P_{s0}. It can be approximated with sufficient accuracy by a straight line—the slope of which is represented by

$$-(dM_0/dP_s)=(\lambda/1.72)\cdot\varphi$$
$$=0.168(E/S)^{0.5}\cdot b\cdot\varphi. \quad (19)$$

According to the value for L/λ gained from Eq. (18), the main equation for calculation of the compression bearing reads

$$R_1^2=R_2^2=(E\cdot b^2/16.3\cdot S). \quad (20)$$

In the particular case of the combination of a bearing using beams under tensile stress and another one with beams under compression stress, the torque of the total arrangement can be kept zero within certain limits if a suitable choice is made of all dimensions. It is therefore necessary to calculate the variation in the force P_s due to any change in temperature or due to a rotation through the angle φ. In case of temperature change the unit stress s varies for both kinds of bearings as follows:

$$\Delta S=E\cdot(\Delta L/L)=E\cdot c\cdot\Delta t. \quad (21)$$

Δt represents the variation in temperature and c the coefficient of linear expansion for one degree of temperature. If in case of temperature change the torque of the combination of two bearings, one under compression, the other under tension, has to remain zero, the mathematical condition can be written

$$\Delta M_{0t}=-\Delta M_{0c}$$
$$=-0.15(E/S_t)^{0.5}\cdot a_t\cdot b_t^2\cdot\varphi\cdot\Delta S_t$$
$$=-0.168(E/S_c)^{0.5}\cdot a_c\cdot b_c^2\cdot\varphi\cdot\Delta S_c. \quad (22)$$

In the above equation the same modulus of elasticity applies to the same material assumed to be used for both bearings. Subscript t indicates bearings under tensile stress; subscript c indicates bearings under compression stress. For equal changes in temperature of both bearings the above condition reads

$$(b_c^2\cdot a_c/(S_c)^{0.5})/(b_t^2\cdot a_t/(S_t)^{0.5})=0.89. \quad (23)$$

To keep the torque of such a double bearing zero independent of the angle of rotation, the relative variations in length $\Delta L/L$ of the beams of both bearings due to a rotation also have to be equal—otherwise Eq. (23) no longer applies. The total change of length of a single beam due to a rotation through the angle φ (see Fig. 5) can be considered to consist of:

(a) A lengthening caused by the curvature of the beam.
(b) A contracting δ caused by the motion of point II on the circle with the radius R_2.

The sum of both changes results in a contracting which is, in case of the compression bearing,

$$\Delta L/L=-0.079\cdot\varphi^2. \quad (24)$$

Choosing for the tension bearing the ratio $L/\lambda=3.15$, the relative change of the beam length equals the expression of Eq. (24). This is the reason for the choice of that ratio.

Now it is comparatively easy to calculate for both bearings the restoring torque resulting from any rotation φ. For the compression bearing results are as follows:

$$\Delta M_{0c}=-1.33\cdot10^{-2}\cdot a_c\cdot b_c^2\cdot(E^3/S_c)^{0.5}\cdot\varphi^3. \quad (25)$$

The torque of the tension bearing is given by

$$\Delta M_{0t}=1.19\cdot10^{-2}\cdot a_t\cdot b_t^2\cdot(E^3/S_t)^{0.5}\cdot\varphi^3. \quad (26)$$

The torque of an arrangement of two bearings obeying Eq. (23) can be kept zero in this way independent of temperature and of the angle φ.

Moreover, there remains the fact that these qualities can also be achieved by a combination of different kinds of bearings under compression stress. One of the bearings is constructed as a normal compression bearing as described above. The other compression bearing is designed so that the compression force P_s corresponds to the second smallest value of solution for L/λ in Eq. (18). This means that with decreasing compression stress the restoring torque of that bearing becomes negative, but with increasing stress, positive. It is obvious that within certain limits, if there is an equal change in stress in both bearings, a compensation of the resulting torques can take place.

Finally, it is worth while to know the effect of radial and axial forces resulting from the weight of the body supported by the bearing. In both radial and axial directions considerable forces can be allowed without any disturbance of the bearing itself. In the case of radial forces it is assumed that the beams of both bearings, seen in axial direction, coincide. In that case, any radial force affecting both bearings in equal parts also produces in corresponding beams the same additional force. Provided that the cross sections of the beams are alike, such a radial force also produces the same unit stress ΔS. The resulting torques of both single bearings compensate each other if Eq. (23) is

satisfied. The torque of the total bearing arrangement remains zero. A motion of the axis of the bearing in the direction of the radial force practically takes place in the same ratio as the change in length of the beams under the influence of that force. This means it is negligible in most cases.

Summarizing, the following conclusions may be drawn. The torque of the total bearing arrangement remains zero if the radial load produces equal unit stress in the corresponding beams of both bearings. In case of equal cross sections of the beams ($a_t \cdot b_t = a_c \cdot b_c$) the radial force must be divided into two equal parts acting on both bearings. Concerning the forces due to accelerations (weight), the center of gravity of the supported body must be situated in the middle between both bearings.

The axial forces have to be counteracted by the bending moments of the beams. Additional compression stress results on the one side of the neutral axis, additional tensile stress on the other side. Obviously the sum of those additional forces does not produce an additional torque in the beam because the torque caused by compression results in a reverse sign from the one caused by tension. Because in most cases the length of the beams is relatively short and the moment of inertia of the cross section ($b \cdot a^3/12$) comparatively high, considerable forces may be admitted in this direction without affecting the bearing.

Finally, an example will show what can be expected of such a bearing. An arrangement consisting of a tension and a compression bearing to support a system of 10 kg of weight has to be calculated. It is based on beams of bronze ($E = 11{,}000$ kg/mm^2). The unit stress in case $P_s = P_{s0}$ is chosen to be 25 kg/mm^2. The cross section of the beams is to be 3 mm^2. Then the resulting force of that beam is 75 kg. The side a_t of the beams of the tension bearing is assumed to be 5 mm; then b_t equals 0.6 mm.

Equation (10) yields for the tension bearing

$$R_1 = 3.64 \text{ mm},$$
$$L_t = 11.45 \text{ mm},$$
$$R_2 = 7.81 \text{ mm}.$$

The assumption of an equal cross section of the beams of the compression bearing yields by Eq. (23) for b_c

$$b_c = 0.89 \cdot b_t = 0.535 \text{ mm}.$$

The other side results in

$$a_c = (a_t/0.89) = 5.62 \text{ mm}.$$

Equation (20) yields for the compression bearing

$$R_1 = R_2 = 2.79 \text{ mm},$$
$$L_c = 5.58 \text{ mm}.$$

An estimation of the moments of the total bearing arrangement, as far as they are caused by inaccuracies of the construction, can be based on Eq. (26). The torque resulting therefrom has to be multiplied by three—the number of beams of each single bearing. A rotation through an angle of 1° results in

$$3\Delta M_{0t} = 7.9 \text{ cm g}.$$

In case of change of temperature Eq. (11) can be transformed to

$$3\Delta M_{0t} = 0.45 \cdot (E^3/S_t)^{0.5} \cdot b_t^2 \cdot a_t \cdot c \cdot \Delta t \cdot \varphi. \quad (27)$$

Then a variation in temperature of 5°C yields for beams of bronze and an angle of 1°, according to Eq. (27),

$$3\Delta M_{0t} = 28.6 \text{ cm g}.$$

Assuming that as a result of those inaccuracies 3 percent of the above-mentioned torques are not compensated for in the most unfavorable case, a variation in temperature of 5°C and an angle of rotation of 1° yield a torque of 1.1 cm g.

On Supersonic Flow Patterns

M. M. Munk
Naval Ordnance Laboratory, White Oak, Maryland
(Received May 3, 1948)

For a definite Mach number, given by $M^2=4/(3-\gamma)$, all Mach lines in a steady plane two-dimensional potential flow of a perfect gas with expansion exponent γ have zero curvature. The use and significance of this relation are discussed, and the hypothetical expansion law consistent with straight Mach lines is deduced.

INTRODUCTION

THIS paper deals with the streamline pattern and Mach-line pattern of perfect gases executing a steady plane two-dimensional potential flow. All Mach lines connect, then, points associated with predetermined families of velocity vectors. The graphical method for the construction of the patterns, first pointed out by Prandtl and Busemann,[1] makes use of this. This method has by now been frequently described and can be considered as well-known. The graphical method can be replaced by equivalent numerical methods.

Any such graphical or numerical work is laborious. It is eased by giving attention to the general features of the patterns studied. It is particularly helpful to give full attention to the shape of the Mach lines. The Mach-line pattern is complicated near the sonic velocity region, but it soon reaches considerable smoothness and regularity; the Mach lines even straighten out.

To investigate that closer, let us introduce the angle of flow direction θ, the angle of Mach-line direction λ, and the Mach angle μ, $\sin\mu=1/M$, so that

$$\lambda=\theta\pm\mu. \tag{1}$$

We prefer the equation

$$\lambda=\theta-\mu. \tag{2}$$

Equation (2) can be made to serve both Mach lines at any point by choosing the positive sense of the angle μ as clockwise or counterclockwise.

CHARACTERISTIC EQUATION

Along any characteristic in a steady plane two-dimensional flow it follows from Meyer[2] (Eq. (17)) that

$$d\rho/\rho=d\theta/(\sin\mu\cos\mu)=M^2d\theta/(M^2-1)^{\frac{1}{2}}. \tag{3}$$

This equation holds for any barotropic fluid (pressure p being a function of density ρ only). For a perfect gas this relation reduces to

$$d\ln p/\gamma=d\theta/(\sin\mu\cos\mu). \tag{4}$$

This equation is written with the reduced variables discussed by Munk and Prim.[3]

Equation (3) expresses the fact that under the conditions stated the velocity increments along any characteristic are normal to the conjugate characteristic. This velocity increment is entirely ascribable to the component of the pressure gradient normal to the conjugate characteristic. Hence

$$dp=cd\bar{w}\rho,$$

where c denotes the velocity of sound, this being the component of the velocity normal to the Mach line, and $d\bar{w}$ the velocity increment normal to the Mach line. Hence the change of flow direction angle is

$$d\theta=d\bar{w}\cos\mu/w. \tag{5}$$

This, in conjunction with $c^2=dp/d\rho$, $M=w/c=1/\sin\mu$, immediately gives Eq. (4).

With a potential flow, Bernoulli's equation $wdw+dp/\rho=0$ holds throughout the flow, not merely along any individual streamline. This serves then to establish a universal relation between the velocity and the Mach number or the Mach angle, and to eliminate the pressure or density from the characteristic Eqs. (3) or (4). For a perfect gas with the ratio γ of the specific heats, there results

$$\begin{aligned} d\log\rho&=d\mu\cot\mu/((\gamma+1)/2+\sin^2\mu),\\ d\theta&=d\mu\cos^2\mu/((\gamma-1)/2+\sin^2\mu). \end{aligned} \tag{6}$$

CHANGE OF MACH DIRECTION ALONG THE MACH LINE

We are interested in the change of the angle λ along the Mach line. The differential of that angle follows directly from (6):

$$d\lambda=d\theta-d\mu=d\mu(\cos^2\mu-\sin^2\mu-\tfrac{1}{2}(\gamma-1))/(\sin^2\mu+\tfrac{1}{2}(\gamma-1)) \tag{7}$$

$$\begin{aligned} d\lambda&=d\mu(M^2(3-\gamma)-4)/(M^2(\gamma-1)+2),\\ d\lambda&=d\theta(M^2(3-\gamma)-4)/(2(M^2-1)). \end{aligned} \tag{8}$$

[1] L. Prandtl and A. Busemann, *Naeherungsverfahren zur zeichnerischen Ermittelung von ebenen Stroemungen mit Ueberschallgeschwindigkeit* (Stodola Festschrift, Zuerich and Leipzig, 1929), pp. 499–509.

[2] R. E. Meyer, "The method of characteristics for problems of compressible flow involving two independent variables," Presented at the 6th International Congress of Applied Mechanics, Paris (1946).

[3] M. M. Munk and R. C. Prim, *On the Canonical Form of the Equations of Steady Motion of a Perfect Gas*, Naval Ordnance Laboratory Memorandum No. 9169, June (1947).

Examining Eq. (8) it appears that the right-hand factor changes sign for

$$M^2=4/(3-\gamma). \tag{9}$$

This is for $M^2=2.5$, with $\gamma=1.4$, corresponding to a Mach angle of 39° 13′ 53″. It follows accordingly that all plane two-dimensional potential flows of a perfect gas exhibit zero curvature of their Mach lines at that special value of the Mach number or Mach angle. The Mach lines will ordinarily have inflection points. This includes all flows, the simple-source flow, the simple vortex flow, and the corner flow (Prandtl-Meyer). The converse also holds. In the corner flow and more general one-parametric flows, the active Mach lines are straight along their entire course, not merely at the special Mach number. But the inactive Mach lines, being intersected by active Mach lines, have points of zero curvature, namely, an inflection point at the special Mach number, and have finite curvature at all other points at finite Mach number. If the pressure gradient at right angles to the conjugate Mach line is not zero, the Mach line will be curved at all finite Mach numbers except at the special Mach number.

REGION OF SPECIAL MACH NUMBER

The magnitude of the velocity for the special Mach angle is $(2(\gamma-1)W_{max}/(\gamma+1)^{\frac{1}{2}}$, or $(1/3)^{\frac{1}{2}}W_{max}$ for $\gamma=1.4$. The square of the special velocity is twice the square of the critical velocity, for all values of γ. The behavior of the Mach lines in the vicinity of that special velocity makes it easier to draw the flow patterns. It does so particularly if the velocity throughout the flow does not change a great deal and is always close to the special Mach number, for then all Mach lines may be replaced by straight lines without undue error. That makes the work very convenient.

Consider, for instance, the case of a parallel flow between two parallel walls $A'A$ and $B'B$ (Fig. 1) with a velocity slightly below the special velocity. At a certain station (points A and B) both walls diverge slightly. The gas will preserve its initial velocity and velocity direction within the triangle ABC. There will be two corner flows within the triangles ACD and BCE. There will be constant and parallel flows within the triangles ADF and BEG, parallel, respectively, to the walls AF and BG. There will be a two-parametric flow within the quasi-diamond $CDKE$.

All Mach lines are straight lines unbroken from wall to wall, even if they extend through different flow regimes. The initial Mach angle must be known. The angle between the initial Mach lines and the inclined walls such as AF furnishes the Mach angle in the newly established parallel flow regions. All Mach lines within the corner flow regions and within the two-parametric flow regions are centered at A and B and are straight.

The corner flow has an angular extension CAD equal to twice the supplement of the wall angle $A'AF$. The change of the flow direction is equal to half the change of the Mach-line direction, and both add up to angle $A'AF$. The change of the Mach angle is equal to the change of the wall angle.

THE SPECIAL MACH NUMBER IN THE HODOGRAPH PLANE

The universal straightening out of the Mach lines in the physical plane at the special Mach number is accompanied by special relations in the hodograph plane at the circle associated with the special Mach number—that is, the circle having twice the area of the circle of sonic velocity. That relation follows from the existence of a Legendre reciprocal potential F in the hodograph plane, the same being subject to a linear differential equation. The gradient of that function F with respect to the velocity vector (u, v) gives the coordinates in the physical plane,

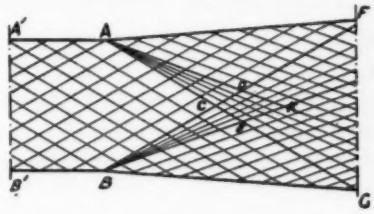


FIG. 1.

$x=\partial F/\partial u$, $y=\partial F/\partial v$. The characteristics in the hodograph plane must be tangent to the direction ellipses. These direction ellipses are discussed in reference 7. Furthermore, the solutions must include functions F constant along one set of characteristics. In other words, the partial differential equation for F must locally degenerate or simplify to the simple wave equation. The special Mach number, Eq. (9), is consistent with an earlier result of the present author,[4] relating to such simplification. The above combination leads to the straightness of the Mach lines, and no other relations would lead broadly to such straightness. R. C. Prim has deduced the differential equation for F in the hodograph plane for the special case $\gamma=1.25$. The mathematical work is laborious even for this special value of γ, and covers a number of pages. Prim obtained the following partial differential equation of the Legendre reciprocal potential F in characteristic coordinates s and t

$$[\tau(\gamma+1)/(\gamma-1)-1]F_{st}+[2\tau(2-\tau(\gamma+1)/(\gamma-1)] \times[F_s+F_t][\tau(\gamma+1)/(\gamma-1)-1][1-\tau]^{\frac{1}{2}}=0.$$

[4] M. M. Munk, "Mathematical theory of gas flow," Aero. Digest (Dec. 1, 1945, and Feb. 1, 1946).

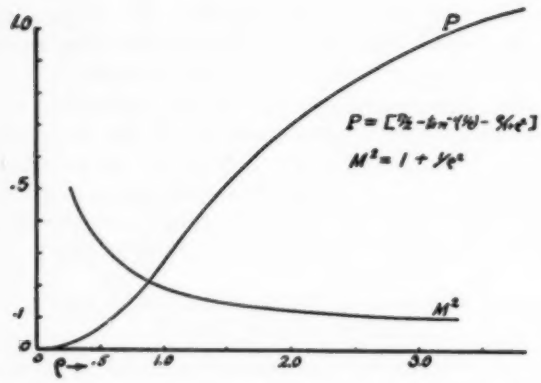


FIG. 2.

Here s and t denote coordinates constant along the characteristics and are the polar angles θ at the intersection or cusp of the characteristics and the sonic circle. Hence $s-t=2\theta$. τ denotes the square of the velocity. According to M. H. Martin, the above equation of Prim holds generally for any value of γ.

BROADER FIELD OF APPLICATION

The substitution of straight lines for the Mach lines at the special Mach number is directly called for by the actual conditions. The Mach lines have actually zero curvature at these points, and hence a straight edge is the best drawing guide with which to trace them.

At Mach numbers other than the special ones, the Mach lines ordinarily have finite curvature, and thus are neither naturally nor well represented by straight lines. Still, the substitution of a train of straight lines to represent a curve has some recommendable features. Equation (9) shows that for any value of M there exists always a value of

$$\gamma=3-(4/M^2), \tag{10}$$

which makes this particular Mach number the special Mach number. The values for γ may become unrealistic, since Eq. (10) yields values of γ between -1 and $+3$. This is not, however, objectionable, since only small portions of the expansion curves are to be used. There are available not only the expansion curves

$$p=k\rho^{\gamma}, \tag{11}$$

but also the modified expansion curves

$$p=k\rho^{\gamma}+\text{const.} \tag{12}$$

It is thus possible to replace the true expansion curve by a train of small portions of modified expansion curves such that everywhere the Mach number (9) of the expansion curve segment and the pressure, density, and the velocity vary continuously. The physical flow is thus divided up into regions—each corresponding to a different expansion law of a perfect gas. The regions are separated by lines of equal velocity and, hence, pressure. They are not separated by characteristics. Hence no sudden change of flow direction θ takes place at the boundaries. The Mach angle changes abruptly. Hence the Mach lines have a constant angular change along each boundary line of constant velocity. Within each region they are straight.

In the diagram of the modified expansion curve p against $1/\rho$ the area $\int(dp/\rho)$ should periodically agree with the corresponding area or integral for the correct expansion curve, in order to keep the error in the pressure-velocity relation as small as possible.

For subsonic flows, H. .Poritsky[5] proposes a similar breaking-up of the expansion curve, leading to simplifications for the computation of the stream function, but not for the Legendre reciprocal potential.

THE HYPOTHETICAL FLUID

Perfect gases give rise to supersonic plane two-dimensional potential flows having locally straight Mach lines. For one special Mach number only is their curvature zero. A more refined analysis of how to replace the gas expansion curve by one having piecewise straight Mach lines calls for the exact expansion curves of a hypothetical fluid giving rise to straight Mach lines. In the hodograph plane, the characteristics must coincide with the curves of constant direction of the conjugate characteristics. That calls for circles of equal radius passing through the center of the hodograph plane. The partial differential equation of the Legendre reciprocal potential (but not of the stream function) in the hodograph plane must become reducible to the simple wave equation.

The discriminant to become constant is

$$\rho^2(M^2-1)=\rho_2^2, \tag{13}$$

ρ_2 denoting the density at the Mach number $\sqrt{2}$.[4] (It also follows from Eq. (4).) This is different from Chaplygin's discriminant $\rho^2/(M^2-1)$, the constancy of which makes the equation for the stream function, but not the equation for the reciprocal potential, reducible.

Combining (13) with Bernoulli's equation leads to the following equation, in which all quantities are made non-dimensional by setting both the density of the gas and the velocity of sound for $M^2=2$, equal to unity, or, in other words, by using these same values as reference quantities, the reference pressure being the product of the reference

[5] H. Poritsky, *An Approximate Method of Integrating the Equation of Compressible Fluid Flow in the Hodograph Plane*, General Electric Company, Data Folder No. 26700.

density and the square of the reference velocity.

$$\rho^2(M^2-1)=1; \quad M^2=(1/\rho^2)+1, \tag{13.1}$$

$$c^2=(4\rho^2/(1+\rho^2)^2); \quad w^2=4/(1+\rho^2), \tag{14}$$

$$M^2=4/(4-w^2), \tag{15}$$

$$\begin{aligned}\rho &= \text{const.}-\tan^{-1}(1/\rho)-(\rho/1+\rho^2)\\ &= \text{const.}+\mu-\sin^2\mu.\end{aligned} \tag{16}$$

In Fig. 2 the expansion curve and the curve for the square of the Mach number have been plotted. The fluid is physically unrealistic. A finite pressure reduces the volume to zero.

The expansion curve has an inflection point, and hence locally constant velocity of sound at $M^2=2$. Equation (9) gives also $M^2=2$ for $\gamma=1$, i.e., constant velocity of sound. At $M^2=1$ the velocity square, by (14), reaches ultimately 4 times the velocity square for $M^2=2$.

The curve in Fig. 2 represents very many curves. The pressure curve may be shifted up and down, and the pressure scale may be changed. The Mach numbers move with the points of the curve, when it is shifted or stretched. In piecing together an actual gas-expansion curve, the same has to be divided into pieces each characterized for the problem in question by its average Mach number. Pieces of the curve of Fig. 2 must be substituted so that the average Mach numbers correspond with each other. Moreover, the velocity ratios at the ends of each interval w_2/w_1 must also agree. This latter condition gives the width of the segment. There then remain two parameters by which to choose the pressure and the pressure interval. The completed compound expansion curve will then represent a continuously varying pressure, density, and velocity, but it will be associated with abrupt changes of the velocity of sound and of the Mach number. The latter will agree sporadically only with the true values.

CONCLUSION

With axial symmetry, with vorticity, or with both, conditions are more complicated—particularly with vorticity, because then the streamlines assume the part of characteristics and must be carried throughout the work. Even then it may be worth while to look out for locally straight Mach lines. If the vorticity is weak, or the axis distant, straightness will occur at Mach numbers not very different from those given by Eq. (9).

The relations discussed throw some light on the significance of Chaplygin's hypothetical gas. It is one thing to replace a curve of locally zero curvature by a straight line, and another to replace a definitely curved line by a straight line. The special Mach number for reducibility of the stream-function equation to Laplace's equation is $M=0$. For small velocities, then, the approximate method of Chaplygin is in its proper place; for any other regime it is less well founded, and the error is much more pronounced.

The Growth and Structure of Thin Metallic Films

HENRY LEVINSTEIN*
Department of Physics, University of Michigan, Ann Arbor, Michigan
(Received June 28, 1948)

The structure of thin metallic films was studied by means of electron diffraction and electron microscopy. Microcrystal size can be correlated with the melting point of the metals, those with high melting points producing continuous films of small unoriented microcrystals while those with low melting points produce large microcrystals oriented preferentially with respect to the substrate. The effects of varying experimental conditions on the film structure of selected metals are reported. The electron beam of the electron microscope affects metals with low melting points most directly, producing melting, sublimation or crystallization. Variations in the velocity of the impinging atom, obtained by passing thermal atoms through a mechanical velocity selector, produce no effect for metals whose vapor is monatomic. Antimony whose vapor is found to consist of polyatomic molecules as well as atoms produces films whose grain size seems to vary with the size of the molecules forming the antimony film. The rate of evaporation merely determines whether zinc or cadmium films may or may not be formed while for antimony films it also determines the type of the films formed. Those formed by rapid evaporation consist of small crystallites while those formed by slow evaporation consist of large amorphous patches. This effect is explained on basis of the mobility of the atoms on the substrate, the mobility time being determined to be less than 2×10^{-8} second. The mobility time and therefore the structure depends upon the type of substrate. The degree of vacuum is critical when the residual vapor reacts with the film being formed, but if no reaction takes place between residual vapor and film, pronounced variation for films formed in a vacuum are produced only when the pressure is above 10^{-3} mm of mercury.

I. INTRODUCTION

THIN metallic films formed by evaporation are characterized by properties which differ considerably from those of the bulk metal. Extensive studies on resistivities, for example, have shown that the resistivity of a thin film is considerably greater than that of bulk metal and that the ratio of resistivity of a film to that of the bulk metal depends upon the thickness and the type of film. Besides the variation in structure between bulk and film, there is considerable variation between films formed under varying experimental conditions. The variables which are most frequently cited as causes for conflicting results are the type of substrate, the temperature during deposition, the velocity and rate at which atoms hit the substrate and the degree of vacuum maintained in the evaporation chamber. Some of these factors have already been investigated. Picard and Duffendack[1] studied the structure of thin films by means of the electron microscope and found that films may be formed more easily on certain substrates than others. Konig[2] observed, by a combination of electron microscopy and electron diffraction, that germanium films which appeared continuous at room temperature, became granular when the substrate was heated during the evaporation. While it is quite tedious to study the effects of all experimental variables on all metals, it is desirable to determine which experimental conditions and which metallic properties are most likely to affect film structure. The most effective tools for such an examination at the present time are a combination of electron microscopy and electron diffraction. Electron diffraction alone cannot give any information on gross structure, except by deduction. While the size and shape of the microcrystals composing the film may generally be estimated from the sharpness of the rings, this method becomes unreliable in two cases. When rings are sharp this technique for measuring crystal size cannot be used to any degree of accuracy because of the uncertainties inherent in the method. When the rings are very diffuse it is often difficult, because of this diffuseness to determine whether the rings are produced by very small microcrystals or by a truly amorphous state.

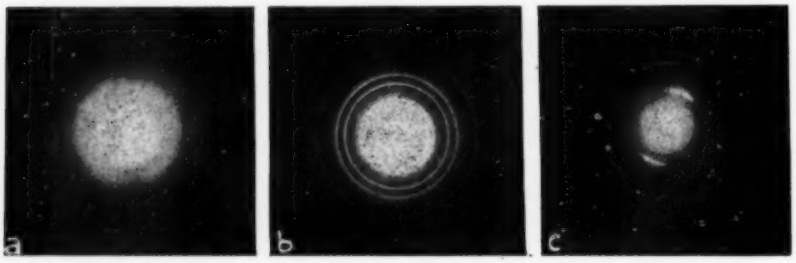


FIG. 1. Electron diffraction patterns of several metal films; a: high melting point metal; b: intermediate melting point metal; c: low melting point metal.

* Now at Syracuse University, Syracuse, New York.
[1] R. G. Picard and O. S. Duffendack, J. App. Phys. 14, 291 (1943).
[2] H. König, Reichsberichte f. Physik 1, 7 (1944).

The electron microscope on the other hand can give no information about films made up of crystallites beyond the resolution of the instrument or about the structure of large patches of atoms. It is the object of this investigation to combine electron microscopy and electron diffraction in a systematic survey of thin metallic films and some factors affecting their structure.

II. THE EFFECT OF THE MELTING POINT OF A METAL ON THE STRUCTURE OF THE FILM

In order to determine which of the physical properties of metals determine film structure most directly, 35 different metals were evaporated on at least 2 occasions in a vacuum chamber at pressures not exceeding 0.01 micron. The films were condensed on a collodion substrate. A range of film thicknesses for electron diffraction studies was obtained by placing several collodion films at different distances from the atom source. For electron microscope examination it was found more convenient to obtain variations in film thickness in such a way that one picture could show the entire range of thickness desired. This was accomplished by placing a razor blade over half the collodion film. Then at the razor edge the penumbra effect due to the extended atom source produced the required variation. It was found that the films studied could be classified in three groups, according to their characteristics.

Group 1: Metals showing diffuse diffraction patterns. Of the metals evaporated the following can be placed in this category: tungsten, tantalum, iridium, columbium, rhodium, germanium, silicon. Of these all but germanium and silicon have a melting point above 1900°C. Figure 1a shows a typical diffraction pattern of a metal in this group. The deposits of all metal films in this group were found to be crystalline, since the ring spacings agreed with those computed from x-ray data on crystalline bulk metal. Had the diffuse rings been caused by an amorphous deposit, different spacings would have been obtained. The crystal size was estimated to be less than 15A. This did not vary appreciably with increased film thickness. As might have been expected from the diffraction pictures, electron microscope pictures showed continuous films for all film thicknesses.

Group 2: Metals showing a sharp line diffraction pattern without preferred orientation. This group includes gold, silver, copper, nickel, iron, cobalt, chromium, manganese, titanium, beryllium, lead, tin, palladium, and platinum. The melting point of these substances with the exception lead and tin are between 600 and 1900 degrees centigrade. It was generally found that for extremely thin films, the diffraction rings were diffuse. They sharpened, however, with increased film thickness. Films with

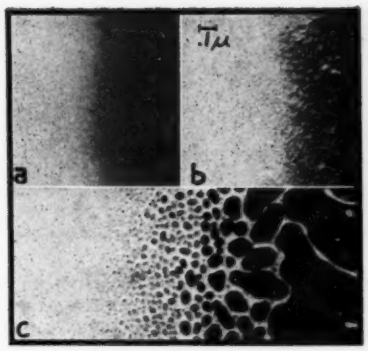


FIG. 2. Electron micrographs showing variations of film structure as a function of thickness for a: high melting point metal; b: intermediate melting point metal; c: low melting point metal.

melting points over 1300°C appeared continuous in the electron microscope. Those with a lower melting point were granular. This subdivision is somewhat artificial, however, since it depends upon the resolving power of the electron microscope and to some extent upon the film thickness. Figure 1b shows a typical diffraction pattern of a film in this group. All patterns in this group were identical with those of Germer[3] and others.

Group 3: Metals producing films whose microcrystals are oriented with respect to the substrate. The following metals are included in this group: antimony, bismuth, tellurium, cadmium, zinc, magnesium, indium, thallium. Thin films were obtained easily for all metals in this group with exception of cadmium and zinc. Methods for forming films of the latter two metals will be discussed later. Without exception these substances have melting points below 650°C. Electron microscope examination of the films generally showed small crystals for thin films growing into large islands for thicker films. Figure 2 shows this behavior for indium compared with the behavior of metals in the other two groups under similar conditions.

Of the metals in group 3 antimony was studied most carefully. Electron diffraction studies by several observers[4,5] have shown that antimony

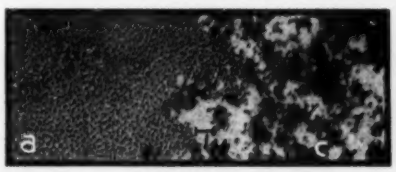


FIG. 3. Partially crystallized antimony film. Region *a* is amorphous. Its electron diffraction pattern shows diffuse rings. Region *c* is crystalline. Its diffraction pattern has sharp rings.

[3] L. H. Germer, Phys. Rev. 56, 78 (1939).
[4] J. A. Prins, Nature 131, 761 (1933).
[5] H. Richter, Physik. Zeits. 19/20, 406 (1943).

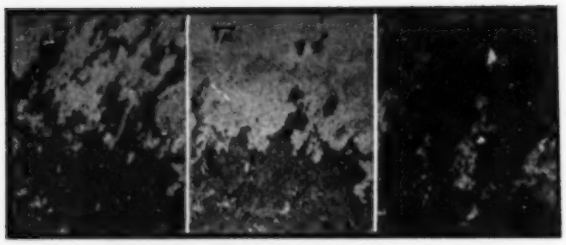


FIG. 4. Apparent variation in the structure of antimony films. The three pictures show the same field of view as the film is turned slightly with respect to the electron beam.

exists in an amorphous as well as a crystalline state. Electron diffraction studies do not in any way indicate the gross structure of antimony films. It is impossible to determine from diffraction pictures whether the amorphous phase is homogeneous or composed of patches of atoms. An electron microscope examination is therefore necessary. Figure 3 shows a partially crystallized film of antimony, the region marked *a* is amorphous, giving diffuse diffraction rings, the region marked *c* is crystalline with considerable preferred orientation. It is interesting to note that the size of the patches in the amorphous region is identical with that in the crystalline region. In the electron microscope the regions differ only in the degree of transparency. The patches in the amorphous region seem homogeneously semitransparent while those in the crystalline region are either completely transparent or opaque. This variation in density or thickness is found to be nonexistant upon closer examination and appears to be due to the crystallinity and preferred orientation of the small patches in region *c*. The effect is similar to that observed by Hillier and Baker[6] on chromium smoke. A microcrystal appears opaque either when it is so thick that electrons cannot pass through it, or when it is so oriented with respect to the electron beam that electrons after being reflected from a particular set of crystal planes cannot pass through the lens system of the microscope. The latter seems to be the case here. Figure 4 shows this effect even more strikingly. All pictures represent the same field of view. The apparent change was produced by changing the inclination of the specimen with respect to the electron beam between exposures. This effect confirms the deductions by Williams and Wyckoff[7] from stereoscopic pictures of antimony films. For fairly thick crystalline films of antimony several types of preferred orientation were observed. When each microcrystal is oriented with respect to the substrate but not with respect to the other microcrystals, continuous rings of anomalous intensities are obtained. These rings break up into arcs when the film is tilted in the diffraction camera. Frequently when microcrystals touch each other, crystallization of one affects the crystallization of the others and together they form one large antimony crystal or several smaller ones. This is shown by the patterns in Fig. 5. The high degree of preferred orientation may also be observed when a thick film tears. Breaks occur along straight lines, making angles of 127 degrees with each other (Fig. 6).

III. THE EFFECT OF THE ELECTRON BEAM ON THE STRUCTURE OF THIN METALLIC FILMS

One of the inconsistencies in electron diffraction studies is produced by the heating effect of the electron beam. This effect is small for substances with high melting and boiling points, but rather noticeable for the others. Substances in group 1 did not change while they were being examined. Of the metals in group 2, gold in particular was observed to change by electron bombardment. Continuous films of gold appeared to granulate when they were exposed to a strong electron beam for a short time. Most metals in group 3 were affected considerably by electron bombardment. Zinc and tellurium sublimed without melting (Fig. 7). In the case of tellurium the thicknesses of the individual crystals did not change during the sublimation process. Holes in the film merely grew bigger, one crystallite evaporating after the other. Zinc crystals on the other hand changed in thickness while the other dimensions remained unaltered. The residue observed is probably an oxide shell which remained after the zinc had evaporated.

Figure 8 shows the effect of the electron beam on thallium films. Thallium melted when first exposed

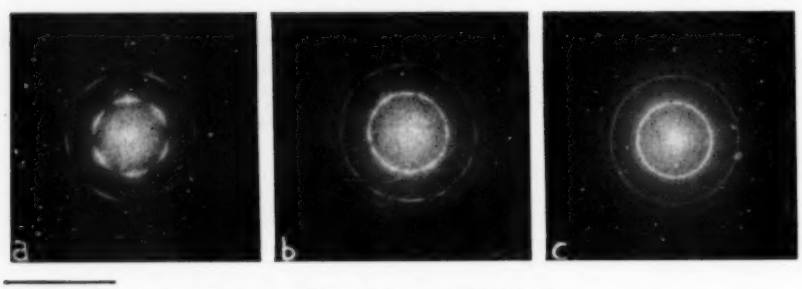


FIG. 5. Varying degrees of preferred orientation in antimony films. In film a the microcrystals are oriented with respect to each other, producing a single crystal. In film b the microcrystals form several smaller crystals. In film c microcrystals are still oriented with respect to the substrate but not with respect to each other.

[6] J. Hillier and R. F. Baker, Phys. Rev. 61, 722 (1942).
[7] R. C. Williams and R. W. G. Wyckoff, J. App. Phys. 15, 712 (1944).

FIG. 6. A broken film of antimony, indicating a high degree of crystallinity.

to the electron beam. The liquid droplets which were formed later evaporated under additional electron bombardment. Indium on the other hand, because of its relatively high boiling point, melted without showing evaporation of the liquid droplets (Fig. 9a). Thin antimony films which were always amorphous, crystallized after some electron bombardment. This process of crystallization revealed itself by the apparent change in density of the previously homogeneous patches composing the film (Fig. 9b). Actually the gross structure of the patches did not change; antimony atoms did not leave the patch, but merely rearranged themselves.

IV. THE EFFECT OF ATOMIC VELOCITY UPON CRYSTAL STRUCTURE

It has been suggested by various observers that the velocity of the impinging atoms might have a considerable effect upon the types of film formed. Miyamoto[8] has developed a theory according to which a potential barrier at a surface would permit only fast atoms to reach the surface while slower atoms would be reflected. Beek, Smith, and Wheeler[9] have suggested that preferred orientation of films, formed by atoms evaporated in a gaseous atmosphere, might be caused by the decreased velocity of the atoms after colliding with the gas atoms. Bateson and Bachmeyer[10] have attributed the poor quality of some evaporated nonreflecting coatings to atoms of low velocity. In order to study the possible effects of atomic velocity on the film structure a mechanical velocity selector was constructed. It consists of a duralumin cylinder into which helical groves are milled (Fig. 10). As the cylinder rotates, atoms pass along a grove without hitting the ridge only if they move in the proper direction with the proper velocity. Atoms can reach a collector directly opposite the source only if they have a velocity equal to the pitch of the helix times the angular velocity of the rotating cylinder. Collectors placed ahead or behind the central collector receive faster or slower atoms, respectively. It is thus possible to obtain a complete velocity distribution during one evaporation. This type of selector has several advantages over the more commonly employed slotted disks or rotating drums.[11,12] It makes it impossible for particles in the vapor beam to reach the collector simply by moving at a sub-multiple of the desired velocity. These particles hit the ridges between the groves. Secondly it makes possible short evaporation times, since the ridges between the groves may be made quite thin, and as great a fraction as 80 percent of the atoms evaporating with the proper velocity may reach the collector. The rotor is placed into a vacuum chamber which is then evacuated to a pressure of 0.002 micron. The metal to be evaporated is placed either in a tungsten filament or an alundum oven in front of the rotor (point A, Fig. 10); the collector consists of several collodion films placed at the rear of the rotor (point B, Fig. 10). In order to maintain a clean vacuum even as the apparatus becomes warm during a long run, it is necessary to avoid the more conventional type motor, with insulated wire or other parts which outgas upon being heated. The rotor is driven by means of magnetic pulses supplied by an electromagnet outside the vacuum. A band painted alternately reflecting and nonreflecting (C in Fig. 10) is pressed around the dural cylinder and an iron bar (D) is attached to the rotor-shaft. Outside the vacuum chamber is a light source and a photoelectric cell, connected by means of an amplifier to the electromagnet. With the rotor turning, pulses of light are reflected from the bulb into the photo-cell producing, in turn, a series of pulses in the electromagnet, thereby driving the rotor, as a motor with a "photoelectric commutator." The selector thus speeds up until the energy fed into

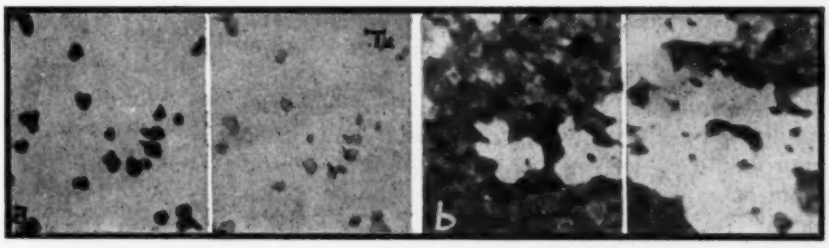


FIG. 7. Sublimation of metals produced by the electron beam of the electron microscope. a: Zinc at the beginning and at the end of an examination; b: Tellurium at the beginning and at the end of an examination.

[8] S. Miyamoto, Trans. Faraday Soc. **29**, 794 (1933).
[9] O. Beek, A. E. Smith, and A. Wheeler, Proc. Royal Soc. **A177**, 62 (1940).
[10] S. Bateson and A. J. Bachmeyer, Nature **158**, 133 (1946).
[11] J. A. Eldridge, Phys. Rev. **30**, 931 (1927).
[12] I. P. Zartman, Phys. Rev. **37**, 383 (1931).

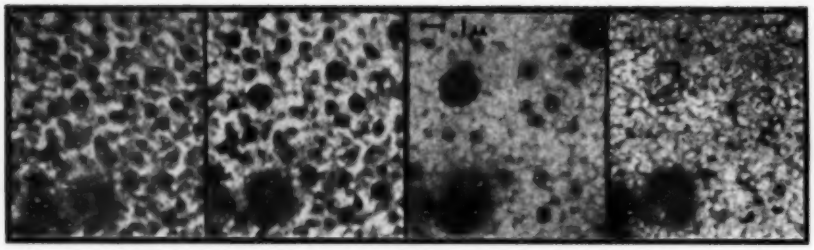


FIG. 8. A sequence of four pictures showing the melting and evaporation process observed during a study of thallium films.

the system equals that dissipated in the bearings. The speed at which the system rotates depends upon the power per pulse supplied by the amplifier, the duration of the pulse as determined by the ratio of the reflecting to the nonreflecting portions of the rotor ring, and the position of the iron bar with respect to the electromagnet when it is energized. The rotor is self-starting, since a small pulse produced by turning on the light source produces sufficient excitation. The speed is determined by means of Lissajous figures on an oscilloscope. Speeds as high as 15,000 r.p.m. are obtained without special effort.

Metals in groups 2 and 3 were studied by means of the velocity selector. Those in group 1 were not included in the study because of their extremely high boiling points and the resulting high velocities. Of the metals in group 2, gold was selected and of those in group 3 antimony and bismuth were chosen.

Gold

The velocity spectrum for gold agreed with that expected on the assumption of a Maxwell-Boltzman distribution for a monatomic gas. The structure of gold films depends upon thickness as has been mentioned previously. In comparing films of equal thickness formed in high and low velocity regions, no variation was observed either in appearance or in crystal structure.

Antimony

Seven evaporations of antimony through the velocity selector were performed. For six of these it was observed that the patches in the high velocity regions were considerably smaller than those in the low velocity regions. Figure 11 shows a typical velocity selection for antimony. Patches in the high as well as the low velocity regions were examined by electron diffraction and found to be amorphous. It was further observed that the distribution curve was not one which might be expected from monatomic antimony vapor with an atomic weight of 122. Assuming a Maxwell-Boltzman distribution, the molecular weight was calculated to be about 300. This indicates that antimony in the vapor phase consists of polyatomic and diatomic molecules and perhaps some atoms. The deposits in the low velocity region of the velocity spectrum are thus produced mainly by polyatomic antimony molecules, while those in the high velocity regions are produced mainly by atoms and by some diatomic molecules. Since the patches in the low velocity regions are formed by fewer but larger molecules than those in the high velocity regions, there are fewer, but larger nuclei of condensation in the low velocity regions than in the high velocity regions with the resulting fewer but larger patches. It seems highly improbable that the variation in size of the patches is due to variation in the velocity of the impinging atoms.

Bismuth

Bismuth is known to exist, in the vapor phase, as a mixture of atoms and diatomic molecules. It may therefore be used to determine whether the variation in deposit structure which was observed for antimony is characteristic of antimony or whether it exists for other metals with a similar type of vapor. Three evaporations of bismuth were performed. The deposits were found to differ considerably from those of antimony, even though the crystal structure of the two metals is the same. Films in the high as well as the low velocity regions were completely oriented, and there appeared no variation in crystal size. The type of vapor is thus not the only factor involved in determining variations in gross structure.

The above experiments seem to indicate that velocity variations within thermal ranges do not affect the structure of the metal films. This may not necessarily be the case when velocities are much higher. High velocities were obtained by charging a 1 μf condensor to 15,000 volts and then discharging it through a 1 or 2 mil wire, thereby exploding it. Under these conditions, however, fragments of metal as well as atoms fly into space. The fragments were found to embed themselves in the substrate. Similar results were obtained by passing a high current through a tungsten filament. Figure 12 shows an electron microscope picture of such a fast evaporation at high temperature. A metal droplet ejected by the filament embedded itself in the previously formed film, leaving a crater. Atoms which were evaporated later cast a

shadow of the droplet, thereby recording the order of events.

V. THE EFFECT OF THE RATE OF EVAPORATION ON THE STRUCTURE OF THIN METALLIC FILMS

From theoretical considerations the rate of evaporation may be expected to have a considerable bearing upon film structure. It has been shown by several observers[13–15] that when the vapor beam density is below a certain critical value, films cannot be formed. This critical density is a function of the substrate temperature and varies widely for different metals. Frankel[16] has explained this by assuming that atoms arriving at the substrate move over the surface and eventually re-evaporate from the substrate. When a collision occurs between two of the atoms which are moving on the surface an atom pair is formed which has a much longer lifetime on the surface than a single atom. Such pairs then act as nuclei of condensation. The critical beam density thus depends upon the liftime of a particular metal atom on a particular surface. Metals in group 3 showed the greatest dependence upon rate of evaporation. Zinc and cadmium films could not be formed at all at low rates of evaporation. When the rate of evaporation was high, films were formed quite readily, but as might have been expected, control of film thickness under these conditions was difficult, and films prepared by this method could not be used for transmission electron diffraction work. For intermediate rates of evaporation isolated zinc and cadmium crystals could be observed similar to those reported by Picard and Duffendack. If evaporation continued long enough, a film was eventually formed with these crystals as nuclei. These effects might be expected for other metals if the substrate temperature were raised, permitting re-evaporation of the atoms, before relatively stable atom pairs might be formed by mobility of the atoms on the surface. Antimony because of its anamolous behavior was studied in detail. While this behavior is not characteristic of many metals at room temperature, it affords an insight into film formation. Figure 13a shows an antimony film evaporated at a low rate (less than 1 atom layer per second), while Fig. 13b shows a film of the same thickness evaporated rapidly.

The large patches of slowly evaporated antimony produce a diffuse diffraction pattern, while the smaller patches of rapidly evaporated antimony produce a sharp ring pattern. This indicates quite clearly that the large patches obtained by slow evaporation are of the non-crystalline form of antimony also observed in very thin films. Five amorphous antimony films were examined periodically over a period of several months. Some crystallization occurred in all films. In two instances there was preferred orientation. In three instances crystallization without orientation was obtained (Fig. 14). No other method was found for obtaining unoriented antimony films by evaporation in a high vacuum.

It appears from these observations that the rate of evaporation affects the crystal structure only indirectly by affecting the grain size of the deposit. When antimony is evaporated rapidly, small patches are formed which at first exhibit no crystalline properties. They are, however, highly unstable in that form, and crystallize rapidly when sufficient atom layers are present to overcome the disordering influence of the substrate. Patches which are formed by slower evaporation, thereby covering a larger substrate area, are much more stable, and crystallize only slowly if at all.

The dependence of particle size on the rate of evaporation may be explained by expanding Frenkel's concept of surface mobility. Two additional assumptions must be made.

1. The number of atoms in motion on the surface is proportional to the number of atoms arriving at the surface per unit time.

2. Atoms or molecules will move over the surface, until they suffer a collision with other atoms or molecules and thereby lose their mobility.

Let us assume that evaporation has been in progress for some time and that there are a certain number of atom patches on the substrate. Newly arriving atoms may either collide with these patches, thereby increasing their size, or collide with each other, thereby forming new patches and increasing

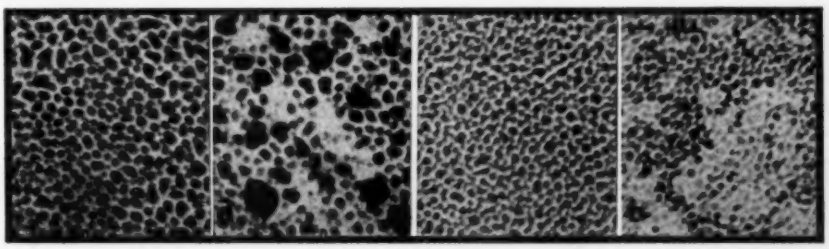


FIG. 9. The effect of the electron beam on the structure of indium and antimony films. a: Two pictures of an indium film. The first was taken just as the film was exposed to the electron beam, the second after some electron bombardment. b: Two pictures of an antimony film. The first shows an amorphous antimony film, the second shows the same film partially crystallized by electron bombardment.

[13] R. W. Wood, Phil. Mag. **32**, 364 (1916).
[14] J. B. Chariton and N. N. Semenow, Zeits. f. Physik **25**, 287 (1924).
[15] J. D. Cockroft, Proc. Roy. Soc. **119**, 295 (1928).
[16] J. Frankel, Zeits. f. Physik **26**, 117 (1924).

FIG. 10. Velocity selector for neutral atoms. A: location of oven. B: location of collodion collectors. C: Rotor band. D: Iron bar used to drive the selector.

the number of the existing patches. If the rate at which atoms arrive at the surface is high, the latter is the case, since the probability for collisions between migrating atoms is greater than that between atoms and already existing patches. Thus many small new nuclei are formed. If atoms arrive at a slow rate, they will more frequently collide with the already existing patches, thereby increasing the patch size. This then results in fewer, but larger patches.

This variation in patch size suggests a method for estimating the length of time atoms remain mobile on the surface. If atoms are evaporated through a rotating slotted shutter disk, they will arrive on the substrate in pulses. The width of a pulse depends upon the width of the slots in the disk and upon the rotational speed. Figure 15A represents an idealized plot of the number of atoms arriving at the substrate *versus* time. Figure 15B shows the same relationship when the motor is turned at half speed. Then more atoms reach the surface per pulse, but there are correspondingly fewer pulses. If the average mobility time of the atoms on the surface is long compared to the pulse time, but much shorter than the time between pulses as shown by line a, then from the previous discussion particle size for case A should be greater than for case B, since there are fewer particles in motion on the surface when the disks are turned faster. Should the mobility time on the other hand be short compared to the pulse time (line b) the number of atoms in motion on the surface would be the same for case A and B, and the particle size would be the same in both cases.

An alternate procedure is available. Rather than changing the rotational velocity of the disk, the slot size of the disks may be increased and the rate of evaporation correspondingly decreased. Then again the number of atoms reaching the surface over a long time interval is the same in both cases. If now the mobility time is long compared with the pulse time (line a, Fig. 16) the patch size will be the same for cases A and B′, since the same number of atoms are in motion on the substrate in both cases. If on the other hand, the mobility time is short compared with the pulse time, patches obtained for case B′ are larger than for case A. The second method may thus serve as a check on the first. If in the first method patch size remains the same, then it must be different in the second, and vice versa. The experiments were performed first with a disk of 2 slots, then with a disk of multiple slots. The disks were driven by the drive used in conjunction with the velocity selector. The shortest pulse time available was 2×10^{-5} second.

The most difficult problem encountered during the experiments consisted of holding the rate of evaporation constant, repeating the evaporation at the same constant rate later, and varying the rate of evaporation between wide limits. The oven found most useful for the experiments consisted of a tightly wound helical tungsten coil of about 1-cm length and $\frac{1}{2}$-cm diameter. One end of the coil was closed with a tantalum shield. About 5 g of antimony in form of a cylindrical slug was surrounded with mica and fitted tightly into the coil. Preliminary experiments were performed to correlate the amount of material evaporated with the current through the coil and the amount of material received on the collector disks. The amount of deposit was determined as follows: A collodion film containing antimony patches was "shadow-cast" with uranium and photographed in the electron microscope. The number of antimony patches and their dimensions could then be determined. Films having an average thickness of as low as 0.1A were measured by this method. It was found that if no atoms were permitted to reach the collector during the initial heating of the filament, the rate of evaporation remained constant until about $\frac{3}{4}$ of the antimony had evaporated from the basket. For the actual experiments it was thus only necessary to set the filament current to an empirically determined value and weigh the antimony sample before and after each run. On several occasions when the amount of antimony evaporated did not agree with the amount expected from the current settings the run was repeated with a new

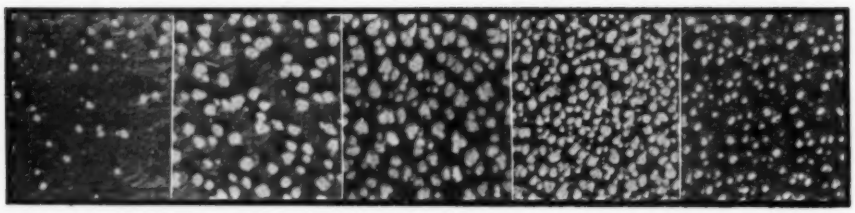


FIG. 11. Antimony patches formed by molecules of varying velocities. The velocity of the molecules forming these patches increased from left to right. The pictures are negatives of antimony films which have been shadow-cast.

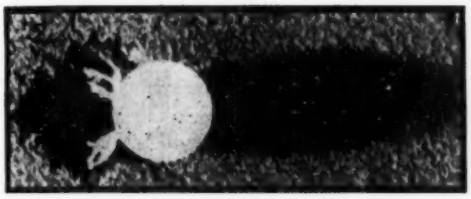


FIG. 12. Result of a rapid evaporation of gold. The small droplet (diameter 1.0 micron) was ejected by the filament forming a crater at its left. It then acted as an obstacle to the atoms evaporating later from the source, thereby producing a shadow at its right. The picture is a negative.

antimony sample. At least two runs were made for evaporation times of $2\frac{1}{2}$, 5 and 10 minutes and a series of pictures was taken of various sections of the collodion films.

The results obtained were entirely consistent. Using the first method the patch size was the same for cases A and B; for the second method larger patches were obtained for case B′ than for case A. This indicated that the mobility time for antimony on collodion was less than 2×10^{-5} second.

The large extent to which mobility of atoms on the substrate affects the structure is more clearly born out, when the substrate is changed. Figure 17a shows antimony on gold. Here the rate of evaporation does not affect particle size, the microcrystals remaining of uniform size for all rates of evaporation. This is probably due to the fact that antimony atoms react with gold and remain at the place they originally hit. Great care must be taken in these experiments against contamination by organic substances, since they seem to affect the surface mobility of the atoms considerably. Figure 17b shows such a surface which had been purposely contaminated by stopcock grease. Patch size varies considerably, and definite conclusions cannot be drawn from such a picture.

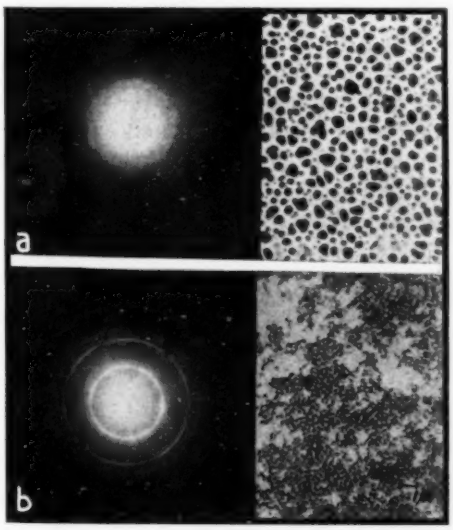


FIG. 13. Antimony films formed by varying rates of evaporation of antimony. a: Electron micrograph and electron diffraction pattern of slowly evaporated antimony which is amorphous; b: Electron micrograph and electron diffraction pattern of rapidly evaporated antimony which is crystalline.

It must be emphasized here that the effects observed for antimony cannot be used to generalize for all metals. The amorphous (noncrystalline) state has never been ovserved for most metals. This could either mean that it does not exist or that crystallization taken place so rapidly that it cannot be observed. The variation in patch size which has been observed for antimony cannot be observed for most other metals at room temperature on collodion because of their low mobility. Raising the substrate temperature might produce the same effect for other metals. However a competing factor would then enter which could not only mask the observations but produce exactly the opposite results. If the difference between the melting point of a metal and the substrate temperature is small, the small patches which are formed may remain in the liquid state for considerable time and several droplets close together might unite to form a large crystallite. If evaporation is rapid there is a high probability that patches are still in a liquid state, when they come close to each other. For slow evaporation, solidification will have taken place long before patches have grown big enough or plentiful enough to touch each other and the crystallites will remain separate. The crystallites formed by rapid evaporation would then be larger than those formed by slow evaporation.

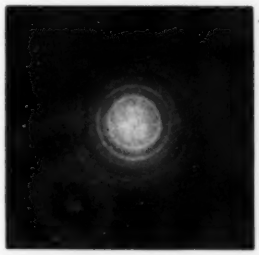


FIG. 14. Unoriented antimony film produced by the crystallization of slowly evaporated antimony.

VI. THE EFFECT OF THE DEGREE OF VACUUM ON THE STRUCTURE OF THIN FILMS

The type of film obtained in a chamber only partially evacuated should be expected to depend upon the type of reaction taking place between the metal evaporated and the residual gas present. The residual gas may affect the film as it is being formed on the substrate or it may affect the individual atoms in transit from source to substrate. What takes place depends upon the pressure of the residual gas, that of the metal vapor and the

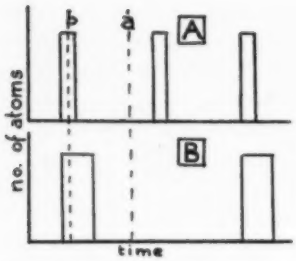


FIG. 15. Idealized plot showing the number of atoms arriving at the film surface *versus* time. A: Rotating slotted disk turning at full speed; B: Rotating disk turning at half speed.

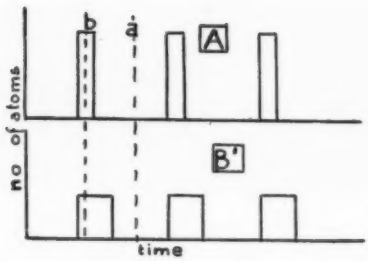


FIG. 16. Idealized plot showing the number of atoms arriving at the film surface as a function of time. This method provides a check on the method indicated in diagram 15. A: Maximum atom beam density, minimum slit opening in shutter disk; B: Atom beam density one half of maximum, slit opening in shutter twice the minimum value.

degree to which chemical reactions may take place between the two. In general it is difficult to determine which of the two produces the observed effect. However, two extreme cases may be considered. When the residual gas is a metallic vapor which may form an intermetallic compound with the metal evaporated, a reaction may take place at the film surface even when the pressure of the residual gas is low. In that case the type of film being condensed and the type of residual gas rather than the pressure of the residual gas are the determining factors. Figure 18 shows such a reaction between zinc and gold. Zinc was evaporated slowly from a filament such that no zinc film could be formed, since the zinc immediately re-evaporated. When gold was evaporated simultaneously, however, a film was formed which showed peculiar configurations. Electron diffraction studies indicated a compound. This behavior is also pronounced when mercury vapor is present in the evaporation chamber. Since the pressure of the residual gas is extremely low it is highly improbable that any reaction between residual gas and metallic vapor could take place while the atoms move from the source to the collector.

The other extreme is the case when a residual gas which does not react with the metallic vapor is used at high pressure. In this case the vapor is not atomic but consists of a cloud of particles which seems to originate near the filament. The gas in this case affects the metallic beam in transit by producing collisions between the individual atoms before they reach the collector, so that the material reaching the collector consists of clumps of atoms. For pressures between 10^{-3} and 10^{-6} mm the results are somewhat inconclusive. An electron microscope examination showed no variation in structure as the pressure was increased to 1 micron. The most striking results were observed when zinc and cadmium were evaporated at a pressure of 0.01 mm. Figure 19, a and b, shows a comparison of zinc films formed in a high vacuum with those formed in a poor vacuum. The films formed in a poor vacuum had a mirror surface and their structure was very uniform. Such films were very suitable for electron diffraction study. A possible explanation of this may be given on basis of Frankel's hypothesis. When zinc atoms arrive at the collector they remain on the surface for only a short time and reevaporate before sufficient other zinc atoms arrive to make film formation possible. When the pressure of the residual gas is increased, collisions between individual metal atoms take place in the vicinity of the filament, thereby producing aggregates of atoms which have a much greater lifetime on the substrate than atoms. These aggregates act as nuclei for other atoms and thus make film formation possible. When the pressure of the residual gas is increased, the number of collisions between metal atoms further increases, resulting in the blackish deposit for most metals. These deposits for most metals appeared very much alike in the electron microscope (Fig. 19c). An interpretation of this structure does not seem possible. It appears, however, that the deposits consist of an array of particles which form near the filament and possess little mobility on the substrate. Electron diffraction study proved useful. Metals which generally showed preferred orientation when evaporated in a high vacuum, showed no orientation when evaporated at pressures of about 0.1 mm. Substances which had produced a sharp line diffraction pattern

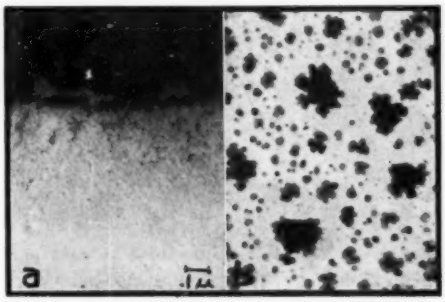


FIG. 17. Dependence of film structure on substrate. a: Antimony evaporated onto a thin layer of gold; b: Antimony evaporated onto a film contaminated with stopcock grease.

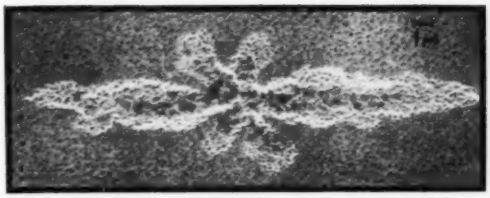


FIG. 18. Film formed by the simultaneous evaporation of zinc and gold.

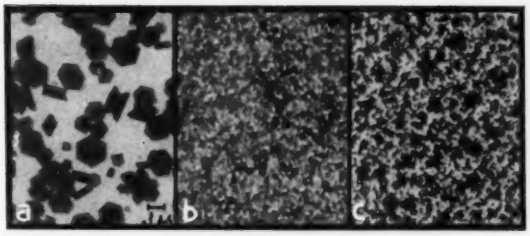


FIG. 19. Dependence of film structure on pressure. a: zinc evaporated at 10^{-5} mm of mercury; b: zinc evaporated at 0.01 mm of mercury; C: zinc evaporated at 0.1 mm of mercury.

when evaporated in a high vacuum showed diffuse rings when evaporated at 0.1 mm. Those which had shown diffuse rings at low pressures continued to show diffuse rings at 0.1 mm. This seems to indicate that for metals in group 1 the particles produced near the filament do not vary considerably in size from those formed on the substrate in a high vacuum. For substances in group 2 particles formed at the filament appear considerably smaller than crystallites formed on the surface. The lack of preferred orientation for particles in group 3 is fairly apparent, and this is understandable, since there seems to be no physical reason as to why small crystallites falling onto a surface should line up in any particular way.

ACKNOWLEDGMENTS

I wish to express my appreciation to Professor H. R. Crane for his many helpful suggestions and his valuable assistance, and to Professor R. C. Williams for his frequent and generous advice in connection with this work.

Use of a Mechanical Harmonic Synthesizer in Electrical Network Analysis*

S. Leroy Brown and Chester M. McKinney
The University of Texas, Austin, Texas
(Received July 27, 1948)

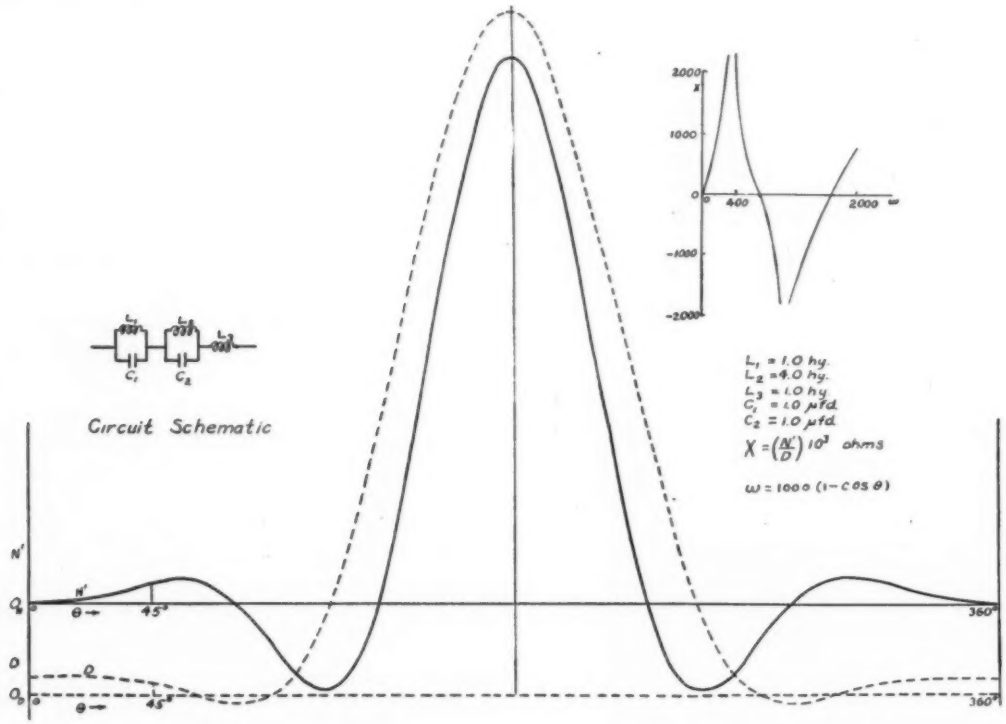
A method is described for determining the variation of the resistance (R) and the reactance (X) of an electrical network from mechanical plots of the numerators and denominators (both functions of the angular frequency ω) occurring in the expressions for R and X. From the expressions, $R=N(\omega)/D(\omega)$ and $X=N'(\omega)/D(\omega)$, mechanical plots of N, N', and D *versus* ω can be obtained with a mechanical harmonic synthesizer, and the values of R or X corresponding to any value of ω may then be determined by taking the ratios of the ordinates of N and D, or N' and D, corresponding to the desired value of ω.

IT is the purpose of this article to demonstrate a mechanical means whereby the impedance of an electrical network is determined from curves that are mechanically graphed for a pre-determined range of frequencies. This aid to the computation of the impedance of an electrical circuit furnishes continuous values for the real and the imaginary components throughout any chosen range of frequency.

The impedance Z of any network with R, L, and C elements can be expressed as:

$$Z=R+jX=[N(\omega)/D(\omega)]+j[N'(\omega)/D(\omega)] \quad (1)$$

where N, N', and D are polynomials in the angular frequency ω. Thus, if the variations of N, N', and D with ω can be graphed, the variations of R and X with ω can be obtained from the ratios of corresponding ordinates read directly from the three curves. The mechanical harmonic synthesizer (consisting of fifteen sine and fifteen cosine harmonic

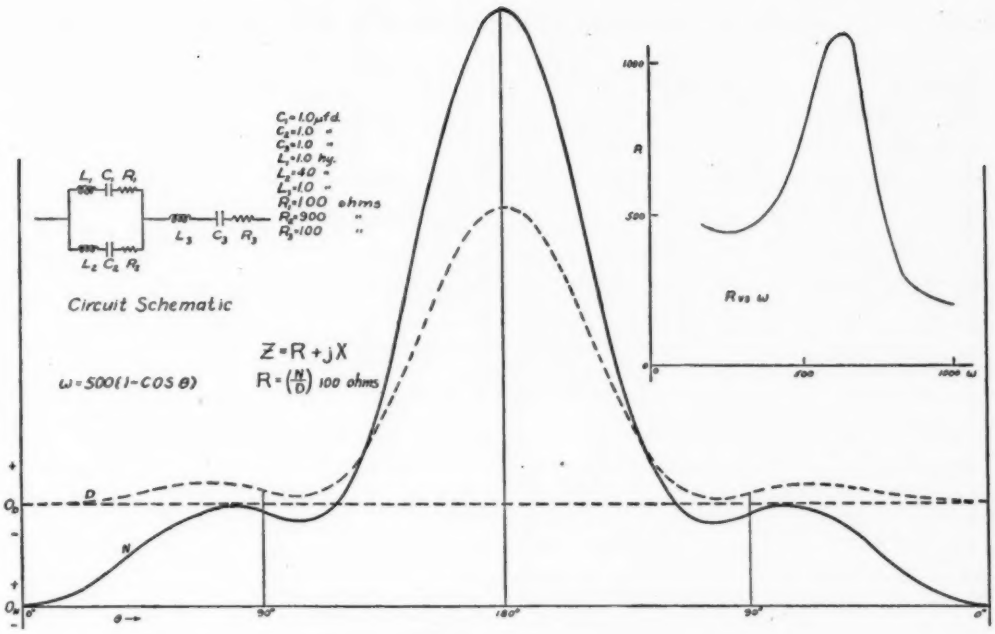


$$X=\frac{4\cdot10^{-12}\omega^5-13\cdot10^{-8}\omega^3+6\omega}{4\cdot10^{-12}\omega^4-5\cdot10^{-8}\omega^2+1}$$

$$=\left(\frac{2.5-4.86\cos\theta+5.25\cos2\theta-4.0\cos3\theta+1.25\cos4\theta-0.13\cos5\theta}{5.5-9.0\cos\theta+5.75\cos2\theta-2\cos3\theta+0.25\cos4\theta}\right)10^3=\left(\frac{N'}{D}\right)10^3$$

FIG. 1. Reactance (X) of a non-dissipative electrical network as determined with a harmonic synthesizer.

* Supported by Office of Naval Research under Contract N6onr-266.



$$R=\frac{5000\times10^{-12}\omega^5-4410\times10^{-8}\omega^3+1400\omega}{25\times10^{-12}\omega^5-19\times10^{-8}\omega^3+4\omega}$$

$$=\left(\frac{5.53-6.85\cos\theta+3.46\cos2\theta-3.02\cos3\theta+0.98\cos4\theta-0.10\cos5\theta}{2.21-3.34\cos\theta+2.29\cos2\theta-1.60\cos3\theta+0.49\cos4\theta-0.05\cos5\theta}\right)10^2=\left(\frac{N}{D}\right)10^2$$

FIG. 2. Resistance (R) of a dissipative electrical network as determined with a harmonic synthesizer.

elements) may be used to graph a polynomial[1] after the polynomial in ω has been transformed into a sum of harmonic functions of an angle θ by a change of variable of the form $\omega = A\cos\theta$ where A is an arbitrary modulus. After the expression for Z has been reduced to the form

$$Z=\frac{\sum C_n\cos n\theta+c}{\sum D_n\cos n\theta+d}+j\frac{\sum C_n'\cos n\theta+c'}{\sum D_n\cos n\theta+d}$$

$$=\frac{N(\theta)}{D(\theta)}+j\frac{N'(\theta)}{D(\theta)} \quad (2)$$

by making the substitution for ω and expanding powers of $\cos\theta$ in terms of functions of multiple angles, the synthesizer may be used to plot N, N', and D and thus, indirectly, to determine the variation of R and X with ω.

The process of obtaining R and X as functions of ω, may be outlined as follows:

(1) Make the change of variable $\omega = A\cos\theta$ or $\omega = A - a\cos\theta$, choosing the numerical values of A and a such that the desired range of ω will be covered as the angle is varied from 0 to 2π radians.

(2) Having made this substitution, reduce the polynomials in $\cos\theta$ to sums of harmonic functions of θ, i.e., $\cos\theta$, $\cos2\theta$, ... $\cos n\theta$ (n not to exceed 15 − the number of cosine elements on the synthesizer).

(3) Set the coefficients of $\cos\theta$, $\cos2\theta$, etc. on the synthesizer as the amplitudes of the 1st, 2nd, etc. cosine harmonics, and vary the angle from 0 to 2π. The graphs of N, N', and D *versus* θ are drawn by the synthesizer and these graphs are easily converted to functions of ω.

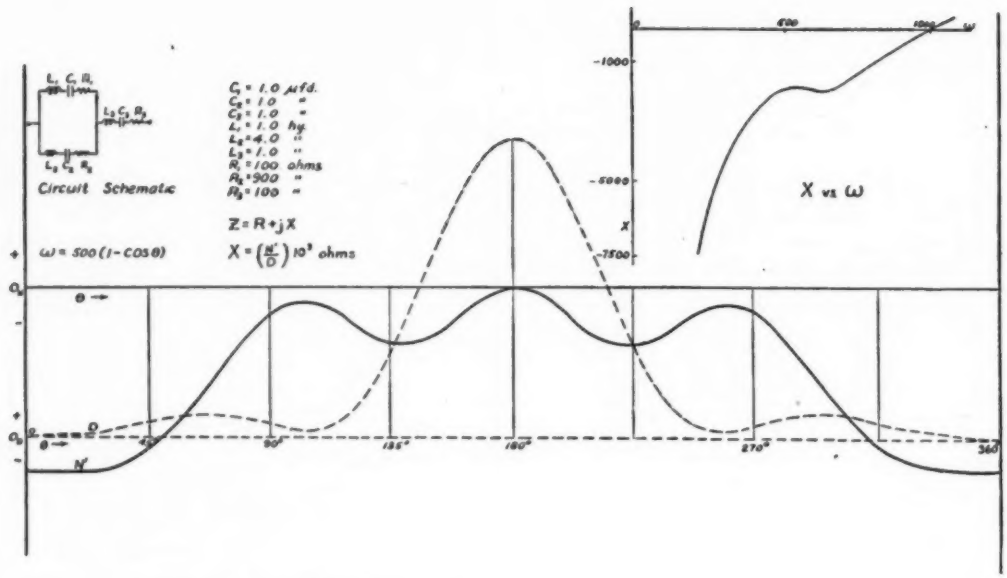
(4) The ordinates are read directly from this curve (the reference axis of which is determined by the constant term in the harmonic form of the polynomial); the ratio of corresponding N and D ordinates and the ratio of N' and D ordinates are the values of R and X, respectively.

The analyses of two types of networks are carried out in some detail to illustrate the use of the synthesizer as outlined above.

I. A NON-DISSIPATIVE NETWORK

The problem is that of plotting the variation of X with ω for the network shown schematically in Fig. 1 for $0<\omega<2000$. The results of steps (1) and (2) above are indicated at the bottom of the figure, which show X first as the ratio of polynomials in ω, and then as the ratio of two sums of harmonic functions. The harmonic expression for X is the result of substituting $\omega = 1000(1-\cos\theta)$—thus causing ω to vary from 0 to 2000—and expressing the resulting powers of $\cos\theta$ in terms of the cosines of multiple angles. If the coefficients of N' are set as amplitudes of the corresponding harmonic elements of the synthesizer, the solid curve (N') is obtained; repetition of this procedure for D gives the dotted curve (D). Values of X are then obtained from corresponding ordinates of the two curves—de-

[1] S. Leroy Brown and Lisle L. Wheeler, J. Frank. Inst. **231**, No. 3 (1941).



$$X=\frac{45\cdot10^{-12}\omega^6-76.15\cdot10^{-9}\omega^4+37.18\omega^2-6\cdot10^6}{25\cdot10^{-12}\omega^6-19\cdot10^{-9}\omega^4+4\omega}$$

$$=\left(\frac{-2.70-2.70\cos\theta-1.12\cos2\theta-0.08\cos3\theta+0.85\cos4\theta-0.26\cos5\theta+0.02\cos6\theta}{2.21-3.34\cos\theta+2.29\cos2\theta-1.60\cos3\theta+0.49\cos4\theta-0.05\cos5\theta}\right)10^3=\left(\frac{N'}{D}\right)10^3$$

FIG. 3. Reactance (X) of a dissipative electrical network as determined with a harmonic synthesizer.

termined from $X=1000(N'/D)$—and from these values, the plot of X *versus* ω shown in Fig. 1 is obtained. It is intended to show, in a later article, a further application of the synthesizer to the analysis of electric wave-filters.

II. A DISSIPATIVE NETWORK

The problem is that of plotting the variation of R and X with ω for the network shown schematically in Fig. 2 and in Fig. 3, for $0<\omega<1000$. The resistance R and the reactance X are first expressed as ratios of polynomials in ω and then as ratios of sums of harmonic functions of an angle. Both of these expressions for R and for X are shown at the bottom of Fig. 2 and Fig. 3, respectively. The harmonic expressions are the results of making the substitution $\omega=500(1-\cos\theta)$, and expanding the resulting powers of $\cos\theta$ in terms of functions of multiple angles. The curves (N) and (D) {as drawn by the synthesizer} and the resulting curve of R *versus* ω—as determined from $R=100(N/D)$—are shown in Fig. 2. Curves (N') and (D), with the resulting curve of X *versus* ω, are shown in Fig. 3.

Theory of Hardness and Measurements Applicable to Contact Problems

ELSE HOLM,* RAGNAR HOLM,* AND ERLE I. SHOBERT II**
Stackpole Carbon Company, St. Marys, Pennsylvania
(Received August 11, 1948)

The theory of the hardness is reviewed and measurements are made with respect to the application of hardness data to contact problems, particularly for the determination of the real contact area. The ball indentation method is recommended, with hardness H defined as the ratio P/S_m between the contact load P and the mouth area S_m of the indentation. This hardness is not a single valued material constant but is a function of the specific depth D of the indentation, where D is the ratio between the actual depth and the radius of curvature of the indentation. The variation of H with geometric and metallurgical conditions is discussed together with some empirical formulae to describe the experimental results. The relation between the hardness and the yield point is discussed. A comparison with Brinell and Rockwell hardness numbers is also made. The indentation which two crossed cylindrical rods produce in each other is found to give nearly the same hardness as the ball indentation tests if D exceeds a certain value.

INTRODUCTION

THE hardness quality of materials is an important property from the standpoint of contact phenomena such as the formation of the contact area, friction, and wear. However, different aspects of what is regarded as hardness are dominating in different applications and in consequence thereof different definitions and methods of measuring exist. The problem is to choose a definition which is advantageous to the description of contact phenomena.

The contact areas between contact members are usually determined by protruding parts of the surfaces with considerable radii of curvature. A similar generation of the contact surface is achieved in the *ball indentation test*. That is the reason why we concentrate upon this method in the present paper. Thereby we define the hardness as an average resistance per unit area against the immerging ball, in agreement with Eugen Meyer.[1] In the Vickers and Rockwell methods there are difficulties which will be discussed below.

We propose to call the hardness defined by the ball indentation method the *contact hardness*.[2]

I. CHARACTERISTICS OF THE INDENTATIONS

The analysis is confined to polycrystalline materials, which have fairly isotropic yielding characteristics. The pressing time was chosen small ($t=4.5$ sec.) and kept fairly constant so that creep and relaxation effects may be neglected in the first approximation.

We now consider how an indentation is generated in a semi-infinite specimen under an immerging ball which is first supposed to be absolutely rigid. If the load is very small the deformation will be purely elastic, defined by the well-known equations of H. Hertz,[3] and will disappear when the ball is taken away. But with a sufficiently large load P the ball may reach the position of the heavy line in Fig. 1.

Displaced material produces a hill around the ball, and the load is carried by the entire surface of the indentation which is indicated by the heavy line in Fig. 1. A part of the deformation under the ball is elastic, and the force due to it is just large enough to produce the counter pressure which carries the load. This elastic deformation vanishes when the ball is removed.*** In consequence thereof the bottom of the indentation goes back the distance f_1 in Fig. 1. The remaining purely plastic deformation is indicated by the dashed line.

The material at the edge of the indentation has reached a high degree of plastic deformation, so that the diameter, which is measured microscopically, is the same as that which existed under the indenting ball except for an elastic displacement which is negligible in practically all cases.†[4] This

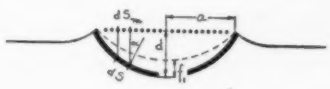


FIG. 1. Ball indentation. Dotted line, diameter of mouth surface. Heavy line, indentation under load. Dashed line, remaining plastic indentation.

* Consultants.
** Technical Director, Rare Metal Contacts and Non-Ferrous Metallurgy.
[1] Eugen Meyer, "Untersuchungen über Härteprüfung und Härte," Physik. Zeits. 9, 66 (1908); also, Mitteilungen über Forschungsarbeiten auf d. Gebiete d. Ingenieurwesens 65–66, 1 (1909).
[2] Ragnar Holm, "Hardness and its influence on wear." Lecture at Conference on Friction and Wear, M.I.T., June, 1948 (in press).
[3] H. Hertz, *Gesammelte Werke 1* (Barth, Leipzig, 1895). The formulas are all listed in: R. J. Roark, *Formulas for Stress and Strain* (McGraw-Hill Book Company, Inc., New York, 1943), Chapter 13.
*** Tabor (see reference 10) has shown experimentally that this deformation is reversible and well counts for the rebound in the scleroscope test.
† The error is definitely negligible in the case of shallow indentations, say for $D<0.1$. See reference 4, Table 8 on p. 20.
[4] H. O'Neill, *The Hardness of Metals and Its Measurement* (The Sherwood Press, Cleveland, Ohio, 1934).

FIG. 2. Illustration of pressure gradient lines xy around a ball indentation.

is of great importance to the reliability of the microscopic measurement of the size of the indentation.

To calculate the average pressure in the indentation, leading to the definition of hardness, we regard an element dS of the surface of the indentation, see Fig. 1. Its normal makes the angle α with the vertical direction of the load. If the pressure on this element is p, the reactive force from it on the ball has the vertical component

$$p \times dS \times \cos\alpha = p \times dS_m,$$

where dS_m is the projection of the element on the horizontal mouth surface of the indentation. Integration gives

$$P = \int p dS_m = \bar{p} S_m, \tag{1}$$

where P is the load, S_m is the mouth area, and $\bar{p}$ is the average pressure as defined by Eq. (1).

Following the custom we regard $\bar{p}$ as the hardness under these circumstances. Thus, designating the *hardness* by H, we have

$$P = HS_m = H\pi a^2, \tag{2}$$

where a is the radius of the mouth area. This definition was first given by E. Meyer.[1]

The trouble with the definition is that H turns out to be very different depending on the shape of the indentations. This effect is well known and the reason will be discussed with reference to Fig. 2.

We regard the immerging ball. The material of the sample yields plastically when the differences between the stresses in different directions are large enough. If σ_1, σ_2, σ_3 are the principal stresses at some point, the condition for a plastic yielding at that point may be written††[5]

$$(\sigma_1-\sigma_2)^2+(\sigma_2-\sigma_3)^2+(\sigma_3-\sigma_1)^2=2Y, \tag{3}$$

where Y is the so-called yield point of the specimen. It is regarded as a constant throughout an isotropic body if no work hardening is generated in it by the deformation. However, the deformations usually cause a work hardening, and Y will vary from point to point and will also be dependent on the direction of the dominating stress.

†† According to v. Mises, see Nadai, reference 9, Chapter 13. An illustration of the content of Eq. (3) is given by E. I. Shobert II in reference 5.

[5] Erle I. Shobert II, comment to the lecture by R. Holm (see reference 2). Conference on Friction and Wear, M.I.T., June, 1948 (in press).

Equation (3) contains stresses. They, of course, stand in a relationship to the generated strains or deformations. The strain region is ruled by the condition of continuity, and the strain is a function $f(x, y, z)$ of the position (x, y, z). However, since Eq. (3) does not contain distance dimensions,

FIG. 3. Indentations with the radii of curvature $-r$, $-r_1$ and $-r_2$ and the same mouth area.

$f(x, y, z)$ will refer to the geometrical shape of the strain or the stress field but not to the size of the field.

Since practically no pressure exists at the free surface of the specimen the pressure must increase along pressure gradient lines yx in Fig. 2 from zero at y up to some value $p(x)$ at the surface of the ball. The pressure $p(x)$ at the point x is the integral of the pressure gradient along yx. This sum is determined by the degree of deformation along yx, but not by the length of yx. A dependence on the number of crystal boundaries along yx is possible, but it usually turns out to be negligible in polycrystalline specimens.†††

The fact that $p(x)$ is defined by the geometric shape of the xy-lines may be explained as a *similarity law:* In the case of similar indentations in a given material the pressure $p(x)$ is the same in homologous points. This means that the hardness H as given by Eq. (2) is the same. A criterion of the similarity between two indentations is that they have the same ratio d_1/r_1 between the depth d_1 and the radius of curvature r_1. We designate d_1/r_1 the *specific depth* and call it D. The similarity law was found empirically by E. Meyer[1] and is well supported by our observations. This means that H is a function of D.

When the ball begins to penetrate, the deformations along the concerned xy-lines are relatively small, and consequently a small hardness H is measured. The deeper the ball penetrates the greater the pressure gradient will be and the larger will be the measured value of H. This means that the hardness defined by Eq. (2) is no single-valued material constant. We shall discuss its significance after presenting the observations. But first it must be shown how the specific depth D may be determined.

II. THE DETERMINATION OF THE FLATTENING OF THE BALL, OF THE ELASTIC RECOVERY OF THE INDENTATION, AND OF THE SPECIFIC DEPTH

Figure 3 illustrates three different indentations with the same radius a of the mouth area. Their

††† The validity of the similarity law is proof that the number of crystal boundaries along the gradient lines xy (Fig. 1) plays a negligible part.

radii of curvature are $-r$, $-r_1$, $-r_2$. The deepest indentation would be obtained with an absolutely rigid ball (radius r). The second one (radius $-r_1$) is obtained with the actual ball. Because of the flattening of the immerged part of the ball, $r_1>r$. The most shallow indentation is the purely plastic one remaining after removal of the ball (see Fig. 1). The corresponding depths may be d, d_1, d_2. Thus the flattening of the ball causes the diminishing of the depth d by the amount

$$f=d-d_1 \tag{4}$$

and the elastic recovery after removal of the ball causes a further decrease of the depth by the amount

$$f_1=d_1-d_2. \tag{5}$$

We confine ourselves to the approximations expressed by the relations

$$d=\frac{a^2}{2r};\quad d_1=\frac{a^2}{2r_1};\quad d_2=\frac{a^2}{2r_2}. \tag{6}$$

A well-known formula by H. Hertz[3] shows that, if two spherical bodies with radii of curvature R_1 and $-R_2$, Young's moduli E_1 and E_2, and Poisson's ratio $\sigma=0.3$, are pressed together by the load P causing a contact area with the radius a, then

$$a^3=\frac{1.37}{2}P\left(\frac{1}{E_1}+\frac{1}{E_2}\right)\left(\frac{1}{R_1}-\frac{1}{R_2}\right)^{-1}. \tag{7}$$

We shall use this equation for two fictitious cases. First we regard the indentation with the radius $-r_1$ as absolutely rigid and assume that the actual ball, with radius r and Young's modulus E_b, is pressed into it until it touches the border of this indentation. This case gives a formula for the flattening of the ball $f=d-d_1$.

Second, we regard an absolutely rigid ball with the radius r_1 pressed against a specimen with initial radius of curvature $-r_2$, and Young's modulus E_s corresponding to the actual specimen, which we consider at this moment to be yielding purely elastically. It is assumed that this ball, when pressed into the specimen, also touches the border of the indentation which has the radius a. Substituting in Eq. (7) we obtain a formula for the elastic recovery: $f_1=d_1-d_2$ of the specimen.

In the first case we have, with $E_1=E_b$, $E_2=\infty$, $R_1=r$ and $R_2=r_1$,

$$a^3=\frac{1.37}{2}\frac{P}{E_b}\left(\frac{1}{r}-\frac{1}{r_1}\right)^{-1},$$

or since

$$\frac{1}{r}-\frac{1}{r_1}=\frac{2f}{a^2}=\frac{f}{rd},$$

we obtain

$$f=\frac{1.37}{2}\frac{Pr}{E_ba^3}d. \tag{8}$$

In the second case we obtain in a similar way, with $E_1=\infty$, $E_2=E_s$, $R_1=r_1$, and $R_2=r_2$,

$$f_1=\frac{1.37}{2}\frac{Pr}{E_sa^3}d, \tag{9}$$

if we define‡

$$\left.\begin{aligned} a_b^3&=\frac{1.37}{2}\frac{Pr}{E_b}\\ \text{and}\qquad a_s^3&=\frac{1.37}{2}\frac{Pr}{E_s}\end{aligned}\right\} \tag{10}$$

in which a_b is the radius of the contact area when the actual ball is pressed with the load P against an absolute rigid plate, and a_s is the radius of the contact area if an equally sized ball of the material of the specimen is pressed with the load P against an absolutely rigid plate.

Hence

$$f=(a_b/a)^3d \tag{11}$$

and

$$f_1=(a_s/a)^3d.$$

We now proceed to the determination of the *specific depth* D, taking into account the flattening of the ball. D is defined by

$$D=d_1/r_1. \tag{13}$$

According to Eq. (6) we have

$$r_1=r(d/d_1),$$

and according to Eqs. (4) and (11)

$$d_1=d-f=d[1-(a_b/a)^3].$$

Hence

$$D=(d_1^2/rd)=\tfrac{1}{2}(a/r)^2[1-(a_b/a)^3]^2, \tag{14}$$

permitting the calculation of D to be made from measured quantities.

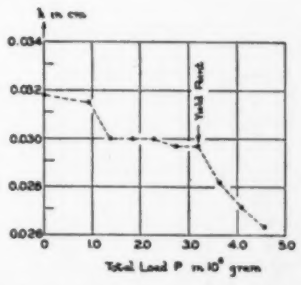


FIG. 4. Illustration of the determination of yield point.

‡ We have calculated with $E_b=2.2\times10^6$ (g/cm²) for the steel ball.

TABLE I. Copper, work hardened and slightly annealed. $F=1.2$.

$2r$ cm	P 10^3 g	a cm	H 10^4 g/cm^2	D 10^{-3}	d_r 10^{-4} cm	H_r 10^4 g/cm^2
0.159	60	0.0774	8.5	17.7	115	8.2
	45	0.0413	8.42	13	81.6	8.4
	30	0.0337	8.4	8.6	57	7.7
	15	0.0240	8.3	4.3	24	7.9
0.318	60	0.0481	8.25	4.3	56.3	8.3
	45	0.0418	8.2	3.3	42.6	8.1
	30	0.0346	8.1	2.2	28.1	7.8
	15	0.0248	7.73	1.1	12.5	7.9
0.635	60	0.0495	7.8	1.1	31.3	7.5
	45	0.0434	7.6	0.84	25.4	6.8
	30	0.0362	7.3	0.57	16.2	6.8
	15	0.026	7.0	0.28	7.4	6.4
1.27	60	0.0525	6.9	0.29	22.1	6.4
	45	0.046	6.7	0.21	14	6.2
	30	0.0382	6.5	0.18	9.1	6.1
	15	0.0273	6.4	0.067	5.2	4.6
	3	0.014	4.8	0.015		

TABLE II. Steel, work hardened and slightly annealed. $F=1.22$.

$2r$ cm	P 10^3 g	a cm	H 10^4 g/cm^2	D 10^{-3}	d_r 10^{-4} cm	H_r 10^4 g/cm^2
0.159	75	0.0356	19	9.2	63	18.6
	45	0.0273	19.1	5.2	34	20
	30	0.0226	18.9	3.5	21.6	20
	3	0.00767	16.3	3.2		
0.318	75	0.0355	18.9	2.07	30.7	19
	45	0.028	18.2	1.24	19.2	17
	30	0.0232	17.8	0.81	12.6	17
	15	0.0167	17.2	0.38	5.4	17
0.635	75	0.0375	17	0.51		
	45	0.0298	16.4	0.3	10.4	16
	15	0.0179	14.9	0.09		
1.27	75	0.0383	16.3	0.09	8.1	18
	45	0.031	14.6	0.06	5.6	15
	30	0.026	14.2	0.03	3.8	14
	15	0.02	11.9	0.019	2.6	9

We have made some measurements on the contact between two *crossed rods* of the same material and both with the same radius r. In this case the contact surface is relatively flat compared with the indentation under a ball. It is circular with nearly the same radius a as a ball indentor with the radius r produces in a plane specimen with the same load. If the axes of the rods have approached each other by δ at the generation of the contact surface, the ratio $\delta/2r$ here plays the same part as D above. Consequently, we have for the rods

$$D=\tfrac{1}{2}(a/r)^2 \tag{15}$$

without any correction. This is based on the consideration that the ratio between the distance d which the material has yielded in the center of the indentation and the mouth radius of the indentation is a measure of its geometrical deformation. If this ratio is the same in the case of a ball indentation and the case of crossed rods, the D's as defined by Eqs. (13) and (15) are the same if we neglect the influence of the hill which has formed a part of the indentation.

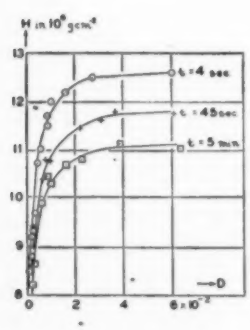


FIG. 5. Influence of the stress time t on contact hardness H.

III. RELATED METHODS FOR DEFINING AND MEASURING HARDNESS

Johan August Brinell[6] introduced his famous ball indentation testing method in 1900. Instead of using Eq. (2), he found that if in Eq. (2) πa^2 be replaced by the curved surface S of the indentation, the hardness number as a function of the load possesses a maximum. In order to obtain a simple method for practical measurements he chose the load and the ball diameter so as to determine a point near the maximum.

With the symbols used above we have

$$S=2\pi r_1 d_1=2\pi r_1^2[1-(1-(a/r_1)^2)^{\frac{1}{2}}],$$

and according to Eqs. (6) and (13)

$$(a/r_1)^2=2D,$$

formally the same equation as Eq. (15). Hence,

$$S=[(\pi a^2)/D][1-(1-2D)^{\frac{1}{2}}].$$

Consequently, the Brinell hardness number H_b may be written

$$H_b=H[D/1-(1-2D)^{\frac{1}{2}}], \tag{16}$$

if H_b and H refer to the same stress time t. Since our H-values are valid for $t=4.5$ sec. and Brinell uses $t=60$ sec. the right member of Eq. (16) should be multiplied by a factor ζ of about 0.93. Values of H_b according to Eq. (16) without the factor ζ have been plotted in Figs. 6, 8 and 14.

The Vickers tester uses a pyramidal indentor. The indentations are geometrically similar regardless of the depth, and therefore always measure

[6] J. A. Brinell, "Ein Verfahren zur Härtebestimmung nebst einigen Anwendungen desselben," Baumaterialienkunde **5**, 276, 294, 317, 364, 392 and 412 (1900). J. A. Brinell and G. Dillner, "Die Brinellsche Härteprobe und ihre praktische Verwendung," Internationale Verbindung für Materialienprüfung, Brüsseler Kongress **27**, 1 (1906).

the same hardness according to Eq. (2). The apex angle is chosen so as to give results in good agreement with the Brinell hardness number.

The Rockwell hardness tester is also widely used. The numbers read on it are related to the depth of the ball indentation. They do not give a measure of the contact hardness directly, but because of the relation between the depth d and the radius a expressed in Eq. (6) it is possible in principle to convert Rockwell readings to contact hardness. Erle I. Shobert II[7] has earlier deduced an approximate formula. It is similar to Brinell's formula, with the difference that P has been replaced by P_1-P_2. It may be written

$$H=\frac{P_1-P_2}{\pi d_r(2r-d_r)}\approx\frac{P_1-P_2}{2\pi r d_r}, \tag{17}$$

where P_1 is the so-called major load, P_2 is the minor

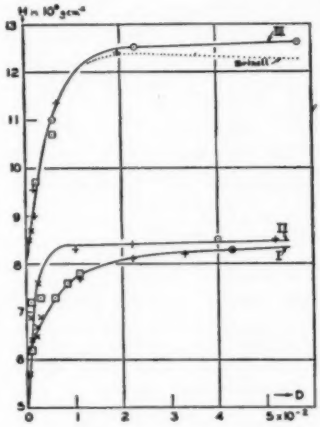


FIG. 6. Copper, work hardened. The units given are for all numbers in horizontal columns.

	I Slightly annealed and slightly polished	*II* Slightly annealed and highly polished	*III* Not annealed	Units
H_0	6.1	5.0	8.1	10^8 g cm^{-2}
H_1	2.1	3.35	4.4	10^8 g cm^{-2}
θ	2.1	1.0	2.5	10^8 g cm^{-2}
D_0	0.7	0.15	0.5	10^{-2}

load in the tester, and d_r is what the tester records as the indentation depth. The flattening of the ball is neglected. The quantity d_r measures the distance between two positions of the ball, loaded with the minor load, taken before and after the generation of the indentation by means of the load P_1. The relation between d_r and the Rockwell hardness number R is

$$d_r=10^{-4}(100-R)\text{ cm}. \tag{18}$$

It should be noted, however, that the Rockwell tester does not account for the influence of the hill

[7] Erle I. Shobert II, "Calculation of electrical contacts under ideal conditions," Proc. A.S.T.M. **46**, 1126 (1946), discussion p. 1145.

TABLE III. Stackpole electrographitic carbon. $F=1.58$.

$2r$ cm	P 10^3 g	a cm	H 10^6 g/cm^2	D 10^{-2}	d_r 10^{-4} cm	H_r 10^{-6} g/cm^2
0.159	15	0.044	2.46	15.3	66.2	2
0.318	30	0.0644	2.3	8.2	68.4	2.35
	15	0.0475	2.1	4.5	32.2	2.1
0.635	45	0.083	2.08	3.4	47.7	2.8
	30	0.0708	1.91	2.5	31.7	2.5
	15	0.0525	1.73	1.37	13.9	2.3

nor for the flattening of the ball. This means that in Eq. (17)

$$\left.\begin{aligned} d_r<d_2\\ r<r_2\end{aligned}\right\}\text{ and }. \tag{19}$$

The exact corrections for Eq. (17), which would provide a conversion from Rockwell hardness numbers, R, to contact hardness, H, are complicated. We limit this discussion to show in the tables that an approximate conversion may be made by means of Eq. (20),

$$H_r=\frac{1}{2\pi r h_r}\left(\frac{P_1}{F}-P_2\right), \tag{20}$$

where F is a quantity which depends on the material of the specimen and the ball, and on the depth of the indentation. For the indentations usually made on the Rockwell superficial hardness tester, F lies between 1.2 and 1.5. Some of the following tables contain observed d_r-values and H_r-numbers calculated according to Eq. (20) with the factor F noted.

IV. EXPERIMENTAL METHODS

The greater part of the indentations for the hardness measurements were produced by means of a

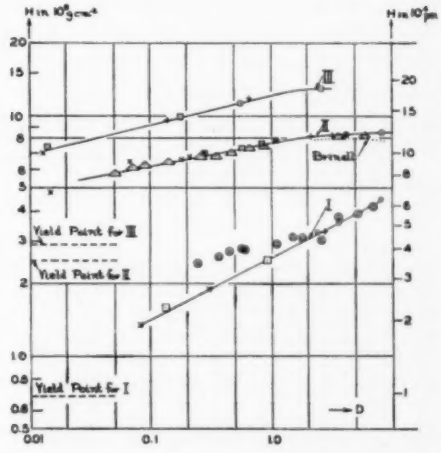


FIG. 7. Copper, work hardened. Curve *I*: completely annealed; *II*: slightly annealed; *III*: not annealed. Double circles refer to measurements with crossed rods. Contact between equal rods. Two different diameters were tested. (A fourth marking, 5.0×10^{-2}, should appear on the abscissa of this figure under the horizontal indicated by D.)

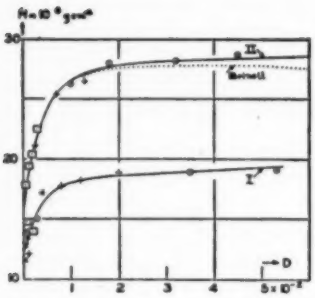


FIG. 8. Steel, work hardened.

	I Slightly annealed	*II* Not annealed	Units
H_0	12.2	16	10^8 g cm^{-2}
H_1	6.1	11.4	10^8 g cm^{-2}
θ	20	20	10^8 g cm^{-2}
D_0	0.4	0.4	10^{-2}

Rockwell superficial hardness tester, normally with a time of stress $t=4.5$ seconds. Some few observations were made with $t=20$, 45, and 300 sec. in order to study the influence of the time.

The measurements with the 0.2-cm diameter ball were made with a Tukon tester with a stress time $t=20$ sec. The corresponding points reduced to $t=4.5$ sec. are marked by triangles in the diagrams. The agreement with the other data is good.

We have chosen the short stress time $t=4.5$ sec. for two reasons: first, to fit the conditions for sliding contacts, and, second, because the Rockwell tester is usually calibrated in this range. Measurements (see Fig. 5) indicate that H does not decrease more than 1 percent during the stress time 4.5 sec. and this is a negligible error.‡‡

The diameter of the indentations was measured microscopically. In the case of elliptic impressions the average diameter was used.

The linear dimensions of the sample must be large compared with the dimensions of the indentations. We did not find determinable deviations in the hardness as long as the thickness of the sample was greater than 3 times the diameter of the indentation. Our diagrams refer to observations on samples with thicknesses of at least 5 times the diameters of the indentations.

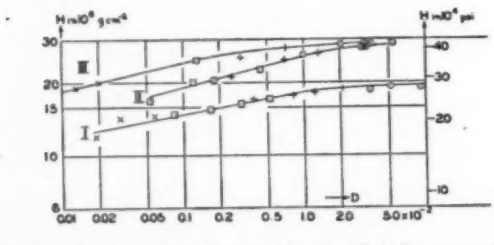


FIG. 9. Steel, work hardened. *I*, slightly annealed; *II*, not annealed, slightly polished; *III*, not annealed, highly polished.

‡‡ Babbit metal is an exception with a corresponding variation of about 3 percent.

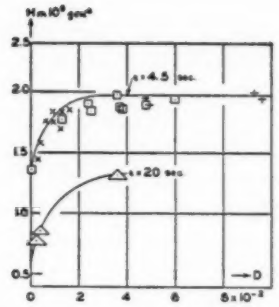


FIG. 10. Babbit with $t=4.5$ sec.

H_0	1.36	10^8 g cm^{-2}
H_1	0.62	10^8 g cm^{-2}
θ	0	
D_0	0.75	10^{-2}

Experiments with *crossed rods* were made with a device of the kind described by R. Holm.[8]

The yield point Y was measured by compression experiments on cylindric specimens with diameters nearly equal to the heights. It is known that such experiments give nearly the same Y as tension experiments, and we do not aim at any great accuracy. Our method is illustrated by Fig. 4. The sample was pressed between parallel surfaces with a force P, the load. Then its height h was measured by means of a dial gauge. The process was repeated with a larger load, then with a still larger, etc.

In Fig. 4 height h is plotted against load P. The first small slope of the curve which connects the plotted points is probably due to the yielding of small protruding parts of the sample. For the metallic specimens the knee at the load P_0 is regarded as indicating the beginning of a general yielding where the height has been diminished 0.1 percent. The contact surface of the sample may be A. Then by definition

$$Y=P_0/A. \tag{21}$$

For the carbon specimen, the yield point was indicated by a breaking or shattering as illustrated in Nadai.[9] To distribute the load evenly over the brittle surfaces the samples were pressed between sheets of soft aluminum. The yield point is represented by the dashed lines in some of the figures.

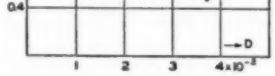


FIG. 11. Tin. Stress time 20 sec. Yield point $Y=0.135\times10^6$ g cm^{-2}.

[8] Ragnar Holm, *Electric Contacts* (H. Geber, Stockholm, 1946), paragraph 8.

[9] A. Nadai, *Plasticity* (McGraw-Hill Book Company, Inc., New York, 1931).

V. EXPERIMENTAL RESULTS OF HARDNESS MEASUREMENTS

The observations are illustrated in the diagrams and partly summarized in Tables I–III. Some of the diagrams have the coordinates H and D, others have log$H$ and log$D$. The symbols of the plotted points refer to the diameters of the balls in the following way:

Symbol	⊙	+	⊡	×	◬	
Diameter	0.159	0.318	0.635	1.27	0.2	cm

The circle within a circle refers to crossed rods.

Figure 5 gives examples of the influence of the stress time t. The reduction to $t=4.5$ sec. has been made on the base of such diagrams. A part of the spread of the plotted points is probably due to uncertainties in the determination of t.

Figures 6 to 14 show that the observed points, although obtained with different size balls, arrange themselves along single curves. This is the statement of the similarity law.

The variation of H with the shallowness of the indentation, which has been discussed above, is clearly established. The H-D-curves for copper and steel drop down suddenly to smaller H-values near the H axis. The slope is less in the case of graphite and still less on commercial tin.

In the logarithmic representation all curves are straightened out, and at least a part of each curve is approximately a straight line represented by the relation

$$H=KD^{\epsilon}, \tag{22}$$

where K and ϵ are constants—for instance, in the case of annealed copper: $K=11.2\times10^{6}$ g cm^{-2} and $\epsilon=0.085$.

Equation (22) is similar to E. Meyer's[1] equation

$$P=k(2a)^{\delta}. \tag{23}$$

However, it must be considered that the logarithmic representation is not sensitive to several of the features which are clearly shown in the H-D-curves.

The following Eq. (24) gives an approximate analytical representation of the H-D-curves. It expresses some of the evident characteristics fairly well. The continuous curves in Figs. 6, 8, 10, 12, 13 and 14 have been drawn according to Eq. (24) with the constants noted under the figures.

$$H=H_0+H_1(1-e^{-D/D_0})+\theta D, \tag{24}$$

where H_0, H_1, D_0, and θ are constants.

The constants have been determined so as to fit the plotted points as well as possible for the range $D>0.001$. Thus H_0 may in some cases be greater than the smallest measured H-values.

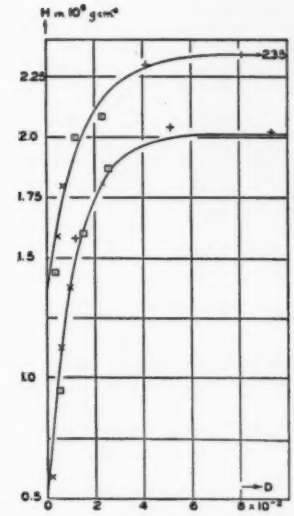


FIG. 12. Stackpole copper-graphite with 45% Cu.

	Curves *I* Contact pressure parallel to molding pressure	*II* Contact pressure perpendicular to molding pressure	Units
H_0	0.54	1.4	10^8 g cm^{-2}
H_1	1.47	0.95	10^8 g cm^{-2}
θ	0	0	
D_0	1.1	1.4	10^{-2}

The hardness H_{cr} measured with crossed rods agrees well with the ball indentation hardness for $D>0.02$, but the crossed rods give higher values for $D<0.01$.

VI. EXPERIMENTAL RESULTS OF COMPRESSION MEASUREMENTS

Using the method which has been described above, measurements have been made on the yield point of some of the materials on which hardness has been investigated. Figure 15 shows the results of yield point measurements on cylindrical samples of annealed iron and of drawn copper. The yield point is plotted as a function of the ratio h/R, where R is the radius and h the height or thickness of the specimen. It is readily seen that the yield point increases as h/R decreases. This increase in yield point, depending upon the shape of the sample, will be a consequence of the varied form (flatness)

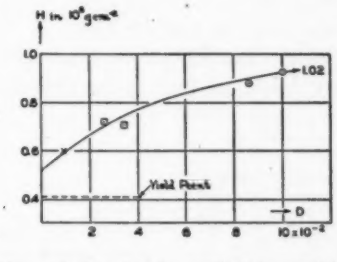


FIG. 13. Stackpole artificial graphite.

H_0	0.35×10^8 g cm^{-2}
H_1	0.36×10^8 g cm^{-2}
θ	0
D_0	1.9×10^{-2}

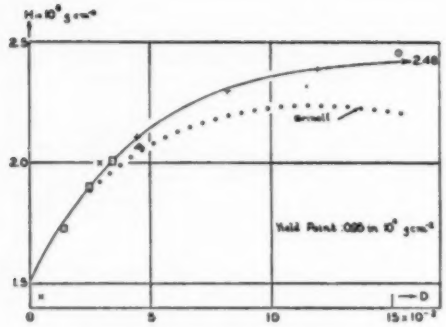


FIG. 14. Stackpole electrographitic carbon. Contact pressure parallel to molding pressure.

H_0	1.51×10^{4} g cm^{-2}
H_1	0.95×10^{4} g cm^{-2}
θ	0
D_0	4.5×10^{-2}

of the field of the generated elastic forces and plastic deformations. It is of the same order as the increase of H in the region $0<D<1$. So we may consider the possibility that this variation of H also may be essentially due to the same phenomena which are illustrated in Fig. 15.

Figure 16 shows some tests made on samples of partly annealed copper in which the load was increased considerably beyond the yield point. It is seen that the values for the average pressure may be made to exceed the hardness values measured by the penetration measurements. This again is probably because of the different geometries, since the material between the plates is bound between narrower boundaries than in the case of the indentors. These tests are described to illustrate the point that hardness as defined by a resistance per unit area is not a simple conception.

VII. GENERAL CONCLUSIONS

The hardness as defined in this paper is a function of the material and of the specific depth of the indentation. The yield point may be regarded practically as a material constant. So the question arises if it would not be sufficient simply to know the yield point. This depends on the purpose for which the information will be used. The yield point is a useful indicator of the elastic limit of the material, and as long as the interest is concentrated upon that limit, hardness measurements are not needed. They have in fact been invented in order

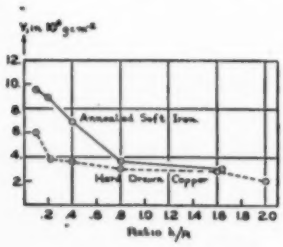


FIG. 15. Variation of yield point Y_1 with the shape of the sample.

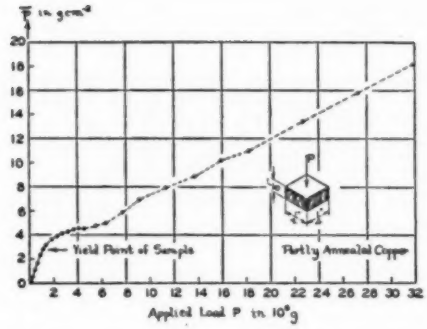


FIG. 16. Average pressure $\bar{p}$ as a function of total load P, particularly the region beyond the yield point Y.

to obtain a simpler practical method than the determination of the yield point.

The Brinell- and Vickers-hardness definitions are bound to a certain specific depth of the indentation. We see from Figs. 6, 8, and 14 that the Brinell hardness corresponds to points in the upper part of the H-D-curves. As mentioned above the Vickers' method defines nearly the same hardness number as the Brinell method. It has been shown (see Tabor[10]) that the Brinell and Vickers hardness numbers are approximately equal to $3Y$ if the work hardening is small. Insofar as this is true the measurement of the yield point and the hardness give equivalent information, and practical use will dictate the simpler method.

Theoretical attempts to explain this ratio 3 exist. Tabor[10] expresses their results in the following way. At the stage where plastic yielding extends over the whole surface of the indentation, the hardness H calculated according to Eq. (2) is nearly equal to $3Y$, if no work hardening occurs. But no singularity in our H-D-curves exists which would indicate where the plastic flow has just spread over the whole surface of the indentation. We observed indentations with a sharp rim and a conspicuous slope at the rim at hardness values as low as $H=1.7Y$ with a stress time of 4.5 sec., cf. Fig. 17. It is likely that the flow begins at the rim of the indentation; but the center is also clearly depressed, and we feel inclined to suppose a general flow also around the shallow indentations. This indicates that in any case the plastic flow all over the indentation is not a complete condition for the relationship $H\approx3Y$.

From Figs. 7, 11, 13, and 14 we read off corresponding ratios of 1.7 up to 3.9 between hardness at $D=0.03$ and yield point. The value of $D=0.03$ is chosen arbitrarily as a point at which there must be considerable plastic deformation. It would appear from Fig. 15, however, that the value of h/R and probably the value of D at which the ratio 3 is

[10] D. Tabor, "A simple theory of static and dynamic hardness," Proc. Roy. Soc. (London) A192, 247 (1948).

reached varies with the material. On the other hand, the effect of work hardening will play a larger part at higher values of D. We believe therefore that the actual curves are a complex of three factors, i.e., the yield point, a geometrical factor which approaches $3Y$, and a work hardening effect, all of which are complicated material properties.

Figure 16 adds a further illustration to the fact that a resistance of the order $3Y$ against an indentor is not a result of work hardening. Since the points plotted in the figure all belong to the limit of the elastic deformation, no work hardening has been produced in this test. Nevertheless, the measurements show a variation of Y_1 from $Y_1=Y$ to $Y_1\approx 3Y$.

Our interest in the hardness was partly motivated by the attempt to find its influence on the mechanical wear in sliding contacts. It is known that hardness plays a part, but it is not clear how it influences wear, or how hardness data should be applied to wear phenomena. It is probable, however, that the wear is dependent on the specific depth of the indentation, which the contact members produce in each other. This would give a certain importance to our investigation of H as a function of D.

Information about the influence of the hardness on the wear is given in references 2 and 8 and we confine ourselves now to some few statements concerning the contact between practically plane surfaces. It is not evident *a priori* what value of the specific depth will be applicable in such contacts. However, measurements of the electric resistance of contacts indicate that the contact surface S usually corresponds to the following equation

$$S=P/H, \tag{25}$$

where H is a quantity of the same order as the contact hardness. It usually corresponds to a hardness number in the range $0.005<D<0.05$ and it is reasonable to suppose that the contact surface is generated essentially by plastic deformation. But exceptions occur. Values of H as small as $H_b/40$, where H_b is the Brinell hardness and where S was determined electrically have been observed on well ground-in contacts.‡‡‡ In such cases, of course, the contact surface was relatively large, and the greater part of it was generated elastically.

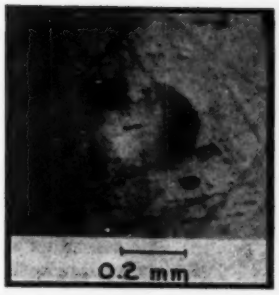


FIG. 17. Indentation in copper with $D=1.5\times10^{-4}$ and $H=4.8\times10^8$ g/cm$^2=1.9Y$. It is evident that the indentation has a finite slope at its rim. The unevenness of the surface is insignificant and does not in any considerable degree influence the form of the indentation.

It would be of great interest to determine to what degree the H-D-curves are due to forces intrinsic in the plastic yielding process and to what degree the so-called work hardening interferes. We do not try to solve this problem now, but only point out that the essential diversity between the H-D-curves for copper or steel on one hand and those for carbon and particularly tin on the other may be of importance for the solution.

ACKNOWLEDGMENTS

We acknowledge with pleasure the assistance of the Westinghouse Research Laboratories for the use of the Tukon Tester, the courtesy of W. E. Kingston of the Sylvania Electric Products Corporation in the preparation of highly polished samples, the interest of R. W. Block of the American Brass Company in supplying samples of drastically rolled copper, and the permission of the Management of the Stackpole Carbon Company to publish the results of these investigations.

‡‡‡ See reference 8, p. 75.

Radiation Resistances of Loaded Antennas*

RICHARD C. RAYMOND AND WAYNE WEBB
Department of Physics, The Pennsylvania State College, State College, Pennsylvania
(Received August 25, 1948)

Short cylindrical antennas have been loaded to resonate at frequencies lower than their natural frequencies by means of dielectric sheaths of cylindrical and conical form. These loadings are compared experimentally with loading by means of metal disks at the ends of the antennas. Resistances at resonance calculated by the Poynting vector method for two assumed current distributions are compared with the experimental results. It is shown that for a given current distribution at resonance, the antenna radiation resistance depends only on the antenna length in wave-lengths. Particular loading methods alter the current distributions. Those which produce more nearly uniform current distributions yield higher radiation resistances.

INTRODUCTION

NUMEROUS theoretical studies of the radiation characteristics of antennas have been published. R. King[1] and his collaborators, Stratton and Chu,[2] and S. A. Schelkunoff[3] have treated the problem from several approaches. In most *practical* cases it has been necessary at some stage in the development to assume a current distribution on the radiator or to assume a field distribution in the vicinity of the conductor because of the difficulty of the boundary value problem. If the major loss of power in a small antenna is regarded as a radiative loss, it is possible to assume that the radiation resistance given by the ratio of the total radiated power to the square of the driving current is the same resistance which is measured at resonance or antiresonance by the conventional standing wave method. In practical use it is often necessary to make an antenna as short as possible without reducing its radiation resistance so far that power cannot be coupled efficiently to and from the antenna over a reasonably wide frequency band. Some form of capacitive or inductive loading is frequently used to reduce the antenna reactance at short lengths and to facilitate operation near the first resonance of the loaded circuit. It is interesting in this connection to determine which loading methods give the higher radiation resistances and permit the shortest antennas.

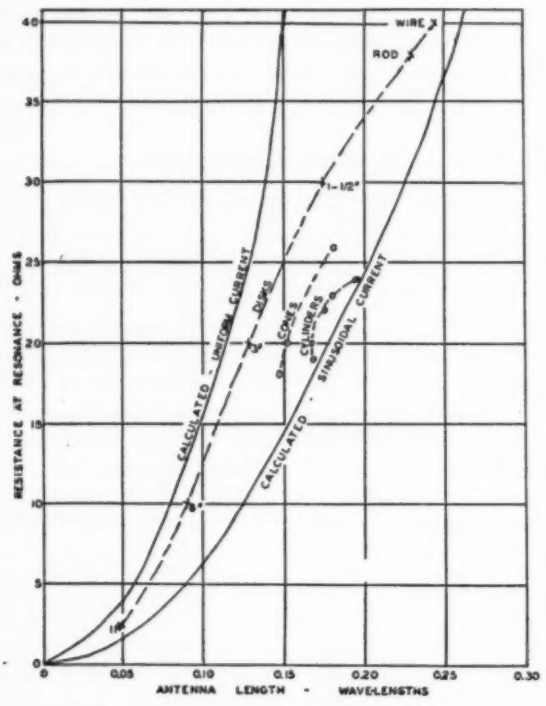


FIG. 1. Resonant resistance for a number of different antennas plotted against antenna length in wave-lengths at resonance. The solid curves are calculated by the method outlined in the text for assumed current distributions. The dashed curves are drawn to connect measured points for three series of antenna models.

In the Poynting vector method of calculating radiation resistance as given by Pierce,[4] the flow of energy through a large surface surrounding the antenna is calculated and divided by the square of the current at the driving point. In the computations below, this method is applied to an antenna which extends in a direction perpendicular to a large ground sheet. An image current is assumed to flow on the other side of the ground sheet, and the surface through which energy escapes is considered to be the hemisphere above the ground sheet. Figure 4 illustrates the arrangement. λ is the wavelength, L the over-all antenna length, and x the coordinate along the antenna. θ is the angle between the antenna axis and the direction in which radiation is considered. If we assume a current, I, given by

$$I = I_0 \cos(\pi x/2L),$$

then the radiation resistance may be calculated

* This work was assisted by the United States Air Force under contract No. W33-038ac7677.

[1] A good review is given by D. Middleton and R. King, J. App. Phys. 17, 273 (1946).

[2] J. A. Stratton and L. J. Chu, J. App. Phys. 12, 230 (1941).

[3] S. A. Schelkunoff, Proc. I.R.E. 29, 493 (1941); J. App. Phys. 15, 54 (1944).

[4] G. W. Pierce, *Electric Oscillations and Electric Waves* (McGraw-Hill Book Company, Inc., New York, 1920).

according to the above method as:

$$R=960\left(\frac{L}{\lambda}\right)^2\int_0^{\pi/2}\frac{\sin^2\theta\cos^2\left(\frac{2\pi L}{\lambda}\cos\theta\right)}{\left[1-\left(\frac{4L}{\lambda}\cos\theta\right)^2\right]^2}d\theta.$$

This integral was carried out numerically to give the curve in Fig. 1 of radiation resistance as a function of antenna length for the assumed sinusoidal distribution.

If, instead of the sinusoidal assumption made above, we assume that the current is uniform along the antenna, the expression becomes

$$R=\frac{15\sin 4\pi L/\lambda}{2\pi L/\lambda}+30\cos 4\pi L/\lambda-60+\frac{120\pi L}{\lambda}\int_0^{4\pi L/\lambda}\frac{\sin t}{t}dt.$$

This expression was evaluated to give the curve of Fig. 1 corresponding to uniform current distribution.

The sinusoidal distribution is the one usually assumed as a first approximation for the case of an unloaded antenna. We might also expect it to hold fairly well in the case of a loading which changes the phase velocity along the antenna but does not alter the current amplitude distribution appreciably. On the other hand, it is probably possible to approach the uniform case by large end loading of a simple antenna.

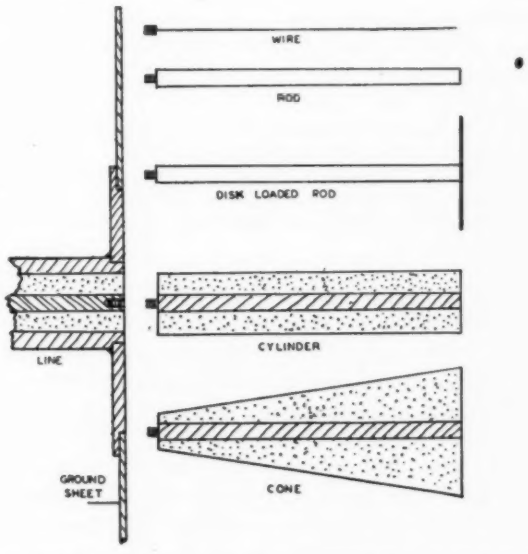


FIG. 2. Outer end of the slotted measuring line and configuration of tested antennas. Each antenna in turn was screwed into the end of the center conductor of the measuring line so that it respresented an extension of the center conductor. Impedances were calculated to the front surface of the ground sheet.

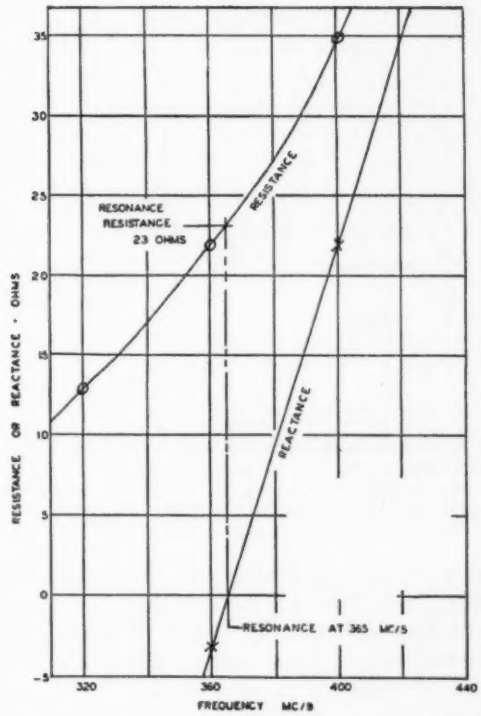


FIG. 3. Plot showing resistance and reactance of antenna loaded with dielectric cylinder in the vicinity of resonance. Resonance was taken as the frequency at which the reactance curve first crossed through zero, and resistance was read from this plot at that frequency.

Along with the theoretical curves of Fig. 1 are plotted measured resistance points for several samples of each of a number of antenna types. These points have been joined by lines where it seems logical that there might be continuous variations between the tested types. All of the tests were made with a conventional slotted transmission line by measuring the amplitude and position of standing wave patterns and calculating the impedances at the surface of the ground screen. Figure 2 is a sketch of the transmission line termination in the ground sheet and of the antenna types used. The transmission line was filled with polystyrene to the level of the ground screen and had an inner conductor $\frac{1}{4}''$ in diameter and an outer conductor whose inner diameter was 0.937". With the exception of the 0.010" wire, all the test antennas were made of $\frac{1}{4}''$ diameter brass rod $5\frac{13}{16}''$ long. They were fitted with threads so that they could be screwed directly into the center conductor of the transmission line.

The cylindrically sheathed antennas were covered with molded layers of dielectric material consisting of polystyrene mixed with varying amounts of titanium dioxide to give dielectric constants of 2.5 to 22.4. The sheaths coincided with the brass rod in length and were of 1" outer diameter. The cones were molded of the same materials $\frac{1}{2}''$ in

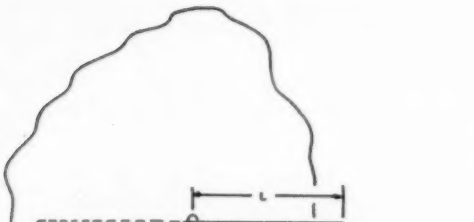


FIG. 4. Sketch showing coordinate system used in calculating resistance by the Poynting vector method. The Poynting vector was calculated at a long distance from the antenna and integrated over the hemisphere in front of ground plane.

diameter at the mounting end, $1\frac{3}{4}''$ in diameter at the outer end. The top loaded antennas were made by soldering successively larger disks to the outer end of the simple brass rod antenna. Diameters of $1\frac{1}{2}''$, $3''$, $6''$, and $11''$ were tested.

Figure 3 is a typical resistance-reactance frequency plot for the cylindrically sheathed antenna of dielectric constant 4.9. It shows resonance at 365 mc/sec. with a resistance of 23 ohms. Resonance was taken for each antenna as the frequency at which the reactance first passed through zero in plots of this type, and the resistances were determined from the same plots.

The experimental results of Fig. 1 agree well in a general way with the theoretical considerations used in deriving the two calculated curves. The measured points for the disk loaded antennas approach the theoretical uniform current curve more closely as the disk diameter increases and approach the sinusoidal current curve more closely as the loading diminishes. The cylindrically loaded antennas fall very near the sinusoidal line, and the antennas loaded with dielectric cones occupy a mid-position between the cylindrically loaded antennas and the disk loaded models. Work reported in another paper[5] leads to a measurement of the current distributions in some of these cases and confirms the conclusions drawn here. From a practical standpoint, neither of the dielectric loading methods is likely to be as satisfactory as the disk loading method.

ACKNOWLEDGMENT

The authors wish to thank Mr. John W. Hoffman, Mr. Donald F. Spooner, and Miss Irene Thomas for skillful assistance with the measurements and calculations reported here.

[5] Wayne Webb and Richard C. Raymond, J. App. Phys. **20**, 330 (1949).

Current Distributions on Some Simple Antennas*

WAYNE WEBB AND RICHARD C. RAYMOND
Department of Physics, The Pennsylvania State College, State College, Pennsylvania
(Received August 25, 1948)

A method is described for measuring the amplitude distribution of high frequency current along some simple types of antennas. Measured curves for some cases are given and compared with some frequently assumed curves. Some of the measured distributions were integrated numerically to determine the driving point resistances. The driving point resistances were measured for geometrically similar antennas in some cases. Deviations from usually assumed curves are significant, but do not bring about large errors in the calculated resistances.

RECENTLY published works dealing with antenna theory[1,2] have shown several methods of attacking the unsolved boundary value problem. In all cases in which the theory has been advanced to the point of calculating the driving point impedance, it has been convenient at some point in the derivation to introduce an arbitrary current distribution along the antenna or to assume an arbitrary electric field distribution near the antenna. The work reported here was undertaken to gain an experimental knowledge of antenna current distributions in a number of cases and to correlate these with measured impedance data.

Because it was desired to conduct the experiments in the region between 250 mc/sec. and 500 mc/sec. to take advantage of existing oscillators, and because the current measuring equipment had to be designed to afford as little interference as possible with the normal operation of the antenna, it was decided to approach the measuring points through the inside of the antenna conductor. The apparatus

* This work was supported in part by the United States Air Force.
[1] A recent review is given by D. Middleton and R. King, J. App. Phys. **17**, 273 (1946).
[2] S. A. Schelkunoff, Proc. I.R.E. **29**, 493 (1941); J. App. Phys. **15**, 54 (1944).

constructed is shown schematically in Fig. 1. It is a short section of coaxial transmission line which is short circuited at one end and flanged at the other for attachment to a large ground screen. The center conductor of the transmission line extends 22.6 cm beyond the ground screen and serves as an element of the antenna under test. The center conductor is made of $\frac{3}{8}''$ O.D. annealed brass tubing and has a $\frac{1}{16}''$ slot throughout its length. It is supported within the line by a molded polystyrene insulator with a keyway opposite the slot in the tubing. A probe entering the line section through the side opposite the keyway is used to feed power to the antenna. Its position is adjustable for tuning purposes.

The current detector is a type 1N22 crystal held in a probe within the center conductor and coupled to the current on the conductor by a small loop which projects through the slot in the tubing. The probe is positioned by a piece of $\frac{1}{4}''$ tubing, connecting with a drive unit and scale outside of the coaxial line. The current measuring probe may be moved continuously from the outer end of the test antenna, through the base transition region and into the coaxial line. The detector output current is read on a spotlight galvanometer. The assembly has been tested for square law operation over the range of current values used, and the crystal current never exceeded 1 microampere in the measurements reported here. The driving oscillator was capable of an output of the order of one watt, and it was therefore not necessary to couple the detector closely to the antenna.

The center conductor was tested as a bare antenna and when modified by two types of loading, as shown in Fig. 2. In one case it was loaded with a cylindrical dielectric sheath of the same diameter as the insulator in the coaxial line and the same length as the antenna. In the other, it was top loaded with a thin metal disk of $4\frac{1}{2}''$ diameter soldered to the outer end of the tube.

In measurement, the oscillator was adjusted to

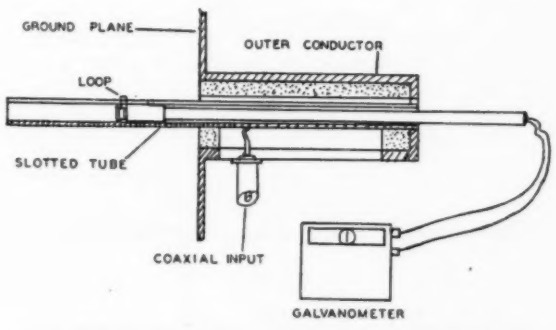


FIG. 1. Schematic diagram of current distribution measuring instrument. The antenna is a hollow slotted tube, and the current detector slides within it. A small loop projecting through the slot couples the detector to the current on the antenna.

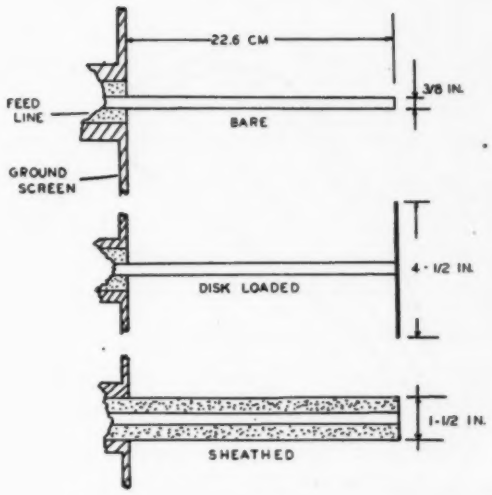


FIG. 2. Dimensions of antennas tested.

a selected frequency, and the antenna system was tuned for proper output. The probe was moved along the antenna, and detector current readings proportional to the square of the high frequency current were recorded every centimeter as galvanometer scale readings. Readings at the outer end of the antenna were discarded because the end of the detector projected and was coupled capacitively to the antenna as well as inductively. Small deviations in the sides of the slot at first introduced errors in the readings near the center of the antenna, but these were finally eliminated by careful machining of the center conductor.

In collateral experiments,[3] antennas similar in geometry to the ones described above were prepared and tested on a conventional slotted line of good construction. The antennas and line termination were made two-thirds the size of the models described here and operated at correspondingly higher frequencies. The antennas were fastened to the open end of a test line similar to that shown in Fig. 2, and impedances were calculated at the surface of the ground screen.

DISCUSSION OF RESULTS

Figures 3, 4, and 5 show typical groups of current distribution curves. From inspection of such groups the resonance and antiresonance frequencies can be approximated quite closely. Table I shows the resonance and antiresonance frequencies determined by inspection from these curves, together with data from the other tests. The frequencies of resonance and antiresonance from the standing wave tests have been divided by 1.5 to correct for the different scale of the antennas tested. The radiation resistance for some of the observed current

[3] R. C. Raymond and W. Webb, J. App. Phys. **20**, 328 (1949).

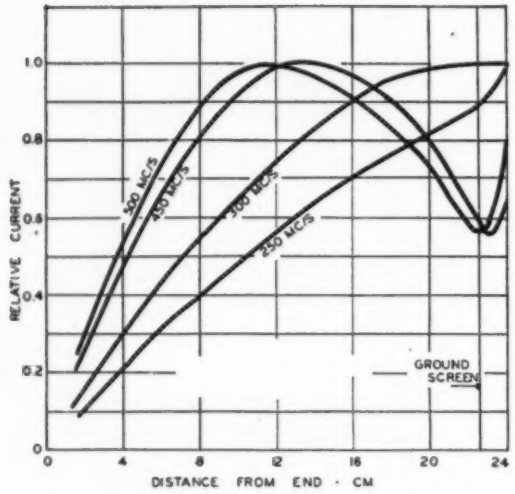


FIG. 3. Current distributions on bare antenna.

distributions on the sheathed antenna were calculated by the well-known method[4] of integrating the Poynting vector over a hemisphere of large radius to obtain the radiated power and dividing the result by the square of the driving point current. Since the current was known as an experimental plot rather than as an analytic function, the integration was carried out numerically. Symmetry was assumed in planes perpendicular to the antenna axis. Twenty-two ordinates were substituted into the trapezoidal rule for the integration along the antenna, and eighteen ordinates were used for the angular integration. The current ordinate at the ground screen was used as the driving point current for determination of the driving point resistance. At 264 mc/sec. the calculated value is 26.7 ohms, as compared with 24 ohms measured for a similarly constructed antenna. Similar calculations at 300 mc/sec. and 420 mc/sec. give the remaining values in Table I. At the antiresonance the calculated resistance is subject to considerable uncertainty because of the small value of the driving point current and its rapid variation with position, near the ground screen. Gap currents at the driving point also render the frequency determination uncertain.

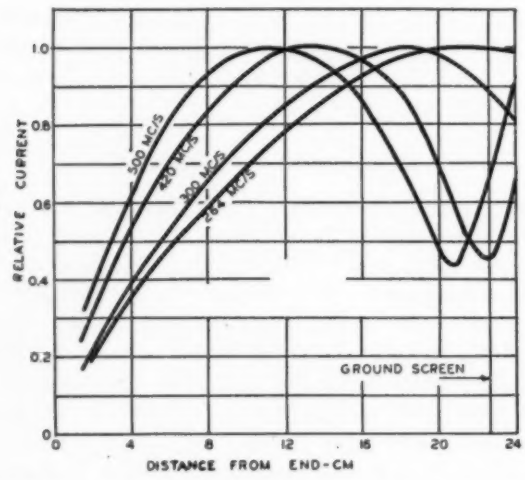


FIG. 4. Current distributions on sheathed antenna.

[4] G. W. Pierce, *Electric Oscillations and Waves* (McGraw-Hill Book Company, Inc., New York, 1920).

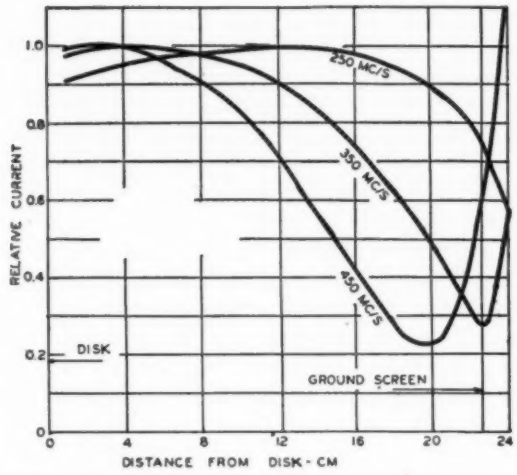


FIG. 5. Current distributions on disk loaded antenna.

A common "handbook" assumption is that the current distribution along an antenna is sinusoidal. Figures 6 and 7 compare the measured distribution on the bare antenna at 300 mc/sec. with the sinusoidal assumption. In Fig. 6, the sine wave was matched at the end of the antenna and at the ground screen. In Fig. 7, the curves were matched at the ground screen and at a point 6 cm from the

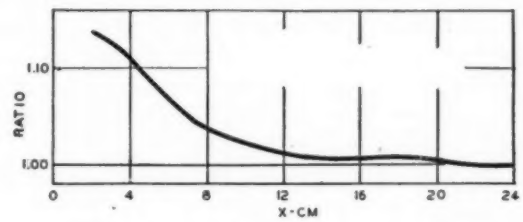


FIG. 6. Ratio of measured current amplitude on bare antenna to assumed sinusoidal distribution given by $\sin(90x/22.6)°$.

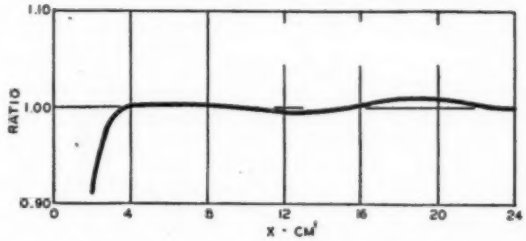


FIG. 7. Ratio of measured current amplitude on bare antenna to assumed sinusoidal distribution given by $\sin(90(x+0.57)/23.17)°$.

TABLE I.

Loading	Scale	Frequency	Resistance		Remarks
None	Full	300 mc/sec.			Resonance
	½	311*mc/sec.	38	ohms	Resonance
Sheath	Full	264 mc/sec.	26.7	ohms calculated	Resonance
	½	265*mc/sec.	24	ohms measured	Resonance
	Full	300 mc/sec.	45.8	ohms calculated	
	½	300*mc/sec.	48	ohms measured	
	Full	420 mc/sec.	329	ohms calculated	Antiresonance
	½		276	ohms measured	Antiresonance
Disk	Full				
	½	173*mc/sec.	20	ohms measured	Resonance
	Full	360 mc/sec.			Antiresonance
	½	376*mc/sec.	630	ohms measured	Antiresonance

* These frequencies are ½ actual measured frequencies to correct for scale differences.

outer end. The latter is seen to give much better agreement and to correspond to the normally used assumption that the antenna acts longer than it actually is.

Near antiresonance, the simple sinusoidal assumption is obviously bad, as it leads to infinite resistance. An assumption selected for comparison is that the current distribution is a standing wave pattern composed of waves traveling in both directions and suffering constant exponential attenuation. Matching this assumption at the end of the antenna, the ground screen, and the maximum current point leads to the dotted curve of Fig. 8. For the bare antenna at 500 mc/sec., it was found that the current function

$$I^2 = I_0^2[\exp(-.05(22.6-x)) + \exp(-.05(22.6+x)) - 2\exp(-.025(45.2))\cos(360°x/22.6)]$$

could be fitted to give the agreement shown.

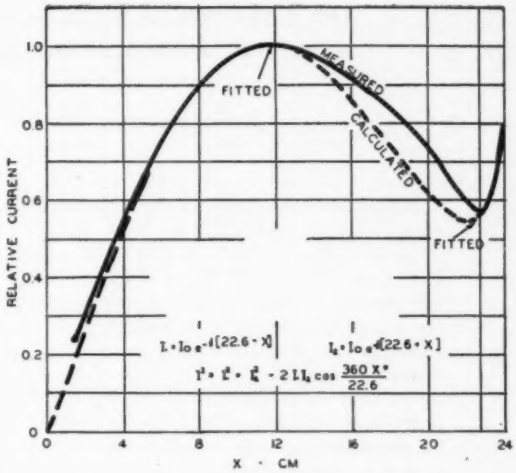


FIG. 8. Comparison of measured current amplitude on bare antenna with assumption based on standing wave in a lossy line.

In conclusion we may say that the measured current distribution is a curve which is not readily fitted by a simple assumption, but that resistances calculated from simple assumptions will be very close to measured values if the curves are fitted in such a way that the large deviations occur where the actual currents are small, as in the case of Fig. 7.

We wish to acknowledge the capable assistance of Miss Irene E. Thomas and Mr. John W. Hoffman who took a great many measurements and did much calculating.

On Theoretical Signal-to-Noise Ratios in F-M Receivers: A Comparison with Amplitude Modulation*

DAVID MIDDLETON
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received August 26, 1948)

Signal-to-noise ratios at the output of an f-m receiver are determined as functions of the input signal-to-noise ratio, clipping level, and i-f and audio filter characteristics when random (fluctuation) noise accompanies the signal to be observed. Both narrow and broad band f-m are examined. Specific calculations are made for sinusoidal frequency modulation, and it is indicated how data for this simple type of signal may be related to the more complex phenomenon of speech. The concept of the signal-to-noise ratio is redefined to describe more accurately the sensing of a signal in noise, since the conventional definition for f-m proves inadequate and misleading, except for strong carriers. Extensive comparisons are made with a-m reception employing a half-wave linear rectifier, and it is found that for moderate and weak (<3 db) signals a-m requires less input signal power against the same noise background than does f-m to achieve the same desired output. This is particularly evident for weak signals. Only when the signal is strong (>10 db) is f-m superior to a-m; here broad band f-m with very heavy limiting is needed. Narrow band f-m at its best (no limiting) is comparable with or inferior to a-m at high signal levels. Limiting proves to be detrimental in narrow-band operation for all cases involving fluctuation noise, whereas heavy limiting is essential for successful broad-band performance. Furthermore, the shape of the i-f response becomes important in the latter, while it is not in the former. A large number of curves illustrating the average and mean-square signal and noise outputs, etc., and signal-to-noise ratios for various conditions of operation are included.

I. DISCUSSION

1. Introduction

THE purpose of the present paper is to determine the signal-to-noise ratio at the output of an f-m receiver, how it depends on such important variables as (a) the input signal-to-noise ratio (a_0), (b) extent of limiting, (c) degree and nature of the modulation, (d) widths and responses of the i-f, limiter, discriminator, and audio filters, etc., and to compare the performance of such a receiver with its analog in a-m reception. In all cases random noise, due to thermal, shot, or suitably random external effects, is assumed to accompany the carrier in the process of demodulation; impulse noise or "static" is not considered here.

The problem is conveniently examined in two parts, according to the relative widths of the i-f and low frequency pass-bands: (a) *narrow band* f-m, for which the maximum deviation D_0 in the carrier's (angular) frequency is comparable with or less than the highest significant audio modulating (angular) frequency ω_a, and (b) *broad band* f-m, which is characterized by a maximum change in carrier frequency that may be much greater than the highest modulating component. A correspondingly wide i-f band is then required for the transmission of the signal. Thus, in the former instance (a) the low frequency noise power after detection may be derived from the entire low frequency noise spectrum; for a given carrier strength spectral shape is relatively unimportant—only the *area* under the spectral curve is required. In the latter (b), however, the power associated with the noise accompanying the signal in the final stage is obtained from a portion of the low frequency spectrum passed by the audio filter, and now spectral shape becomes a significant factor. Similar remarks apply to narrow-band f-m whenever the signal or one of its components is observed in a band narrower than the low frequency noise spectrum following discrimination.

Earlier theoretical work[1,2] has been chiefly restricted to the case of carriers strong relative to the noise. Recently, Blachman[3] has given a treatment of the narrow band f-m problem for all degrees of limiting and input carrier-to-noise ratios, and a discussion of the broad band problem when the carrier is strong ($a_0^2 \gg 1$) and the limiting either extreme or non-existent. Further recent work[4] has

* The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (Office of Naval Research) and the Signal Corps, U. S. Army, under Contract N5ori-76, T. O. 1.

[1] J. R. Carson and T. C. Fry, "Variable frequency electric circuit theory with application to the theory of frequency modulation," Bell Sys. Tech. J. **16**, 513 (1937).

[2] M. C. Crosby, "Frequency modulation noise characteristics," Proc. I.R.E. **25**, 472 (1937).

[3] N. M. Blachman, *The Demodulation of a Frequency-Modulated Carrier and Random Noise by an FM Receiver*, Cruft Laboratory Tech. Report No. 31, March 5, 1948. Some further references to theoretical work are given here. See also Nelson M. Blachman, J. App. Phys. **20**, 38 (1949).

[4] D. Middleton, *The Spectrum of Frequency-Modulated Waves After Reception in Random Noise I.*, Cruft Laboratory Tech. Report No. 33, March 8, 1948, and a paper of the same title to appear in the Quarterly of Applied Mathematics, July, 1949. For a less general analysis, which involves only the case of extreme limiting and an unmodulated carrier, see Sections 7 and 8 of "Statistical properties of a sine wave plus random noise," S. O. Rice, Bell Sys. Tech. J. **27**, 109 (1948). See also an interesting paper by F. L. H. M. Stumpers, Proc. I.R.E. **36**, 1080 (1948).

given us the spectral distribution of the noise from the discriminator for *arbitrary* carrier amplitudes, in the two cases of greatest practical interest, *viz.*, no limiting and extreme or "super"-limiting. With this we may now extend the theory to include a calculation of signal-to-noise ratios for broad band f-m in the threshold region ($a_0^2 \sim 1$). Then, from the new results for broad band f-m derived here and from those applicable to narrow band f-m we can make definite statements regarding the absolute and relative ranges and sensitivities of f-m-*versus*-a-m reception of signals in random noise. Earlier observations on this point have been clouded by the inability to calculate signal-to-noise ratios under any condition other than carriers strong relative to the noise, and, further, by comparison of a-m and f-m receivers functioning under similar circuit design but not necessarily under their respective optimum conditions of operation.[5]

Our reasons for considering narrow band f-m here in the light of Blachman's work stem from the following considerations:

(a) It is possible to derive his results directly and easily by using the probability density function for the envelope and the time derivative of the phase (the present approach), instead of by the more intuitive and less formally direct method based on a vector model of the instantaneous wave and its resolution into components (reference 3, Sections 2 and 3).

(b) Furthermore, an exact expression for the envelope may be used, rather than the approximate form employed in the earlier treatment. The distinction arises because limiting a narrow band wave containing noise produces "spectral zones" distributed about harmonics of the band's central frequency, here the i-f or resonant frequency.[6] Only the band associated with the first harmonic region is passed to enter the discriminator, and it is the envelope of this wave, instead of the limited envelope of the original disturbance, that must be considered. This approximation is not ignorable for moderate or heavy limiting when the signal or the noise contribution is examined separately, although the difference between the two treatments never exceeds 2.1 db ($10 \log_{10} 16/\pi^2$). However, as we shall see presently (Sections 4, 5), the signal-to-noise ratios in narrow band f-m are nearly the same (<0.5 db) in either case.

Finally, a more refined analysis of the signal-to-noise ratio as a criterion of receiver performance is essayed for both types of f-m, since the ratios used in practice leave much to be desired from the point of view of measuring the actual signal as it is perceived in noise. A large number of figures illustrating the theory is included.

2. The Signal-to-Noise Ratio

The conventional definition of the signal-to-noise ratio used in f-m reception is found by taking the ratio of the low frequency discriminator output with carrier and modulation to that with an unmodulated carrier. Such a definition is very convenient operationally, as these quantities are easily measured, but, unfortunately, it does not give a satisfactory criterion of performance for two important reasons:

(a) It does not take into account what is actually perceived as a signal. The signal is merely included with the noise, which in turn contains significant contributions due to the (*signal*$\times$*noise*) noise products generated in the discriminator and/or the limiter.

(b) It fails to consider the effects of *modulation suppression*—whereby a weak signal is suppressed by the stronger noise—i.e., made relatively weaker at the output in relation to the noise than it is at the input. Modulation suppression is observed in all non-linear operations, whether in detectors or discriminators, and is particularly important in determining the maximum range of the device in question. In fact, it can be shown[7] that for input signal-to-noise ratios less than unity *all* half-wave detectors and all (balanced) discriminators behave like a quadratic rectifier in this respect: the output signal is proportional to the *square* of the input (s/n) ratio. On the other hand, for large input carriers the noise is suppressed, and it is only here that the conventional definition of the ratio for f-m may be used, since now the output is essentially the linear sum of signal and noise, albeit the noise is formed from ($s \times n$) modulation products (see reference 4, Eq. (2.40) *et seq.*) and is dependent on the presence of a carrier. In short, the conventional definition fails in the threshold region ($a_0^2 \sim 1$) because it is linear, whereas the process it attempts to describe is non-linear. We must modify the concept accordingly, so that it considers (a) and (b) above.

The definition of the signal-to-noise ratio at the output of the receiver depends on the nature of the signal, which, in turn, predicates the mechanism and mode of presentation. We distinguish two important cases: (I) where the signal itself, e.g., speech or other complex waves, is desired; and (II) where the intelligence to be conveyed consists merely in noting the presence or absence of a signal.

[5] M. G. Nicholson, "Comparison of amplitude and frequency modulation," W. Eng. **24**, 197 (1947). This paper also contains a fairly extensive bibliography.

[6] D. Middleton, "Some general results in the theory of noise through nonlinear devices," Quart. App. Math. **5**, 445 (1948); see, in particular, Section 3 and Fig. 3.

[7] D. Middleton, "Rectification of a sinusoidally modulated carrier in the presence of noise," Proc. I.R.E. **36**, 1467 (1948). See also Cruft Laboratory Tech. Report No. 45, July 1, 1948.

The former requires a band width sufficient to include all significant components, while for the latter we need consider only a suitable harmonic of the modulation. Some possible ways of observing the signal in noise are,[8] *as speech:*

(a) *By earphones*, where the entire wave with its associated noise is passed in the final stage. Phase relations between components must be preserved, since it is the instantaneous envelope that is important, as in the corresponding radar problem.[8] A sound-level meter may be used for sufficiently intense signals.

(We can also note the *presence or absence* of a signal):

(b) *By earphones*, where but one harmonic of the signal is required. Here a narrow audio filter is used to isolate the desired component. The critical band width of the ear[8] may be significant.

(c) By a *meter or recording apparatus:* these respond to d.c. or to the very low frequencies (several cycles/sec. or less) obtained by rectifying the output of the narrow audio filter of method (b), followed by an appropriate low pass filter.

(d) By a *bolometer or thermocouple*, which measures the *area* under the distribution curve of the spectral power density. Knowledge of phase is consequently lost. The bolometer is convenient for power measurements in any given band, whereas the narrow filter or meter of (b) and (c) must be employed for the signal.

To be fully effective as a criterion of performance and comparison with other modes of reception the signal-to-noise ratio at the output in the case of speech should be related to the observer and his psychophysical responses. This is possible when we wish to determine the value of the minimum intelligible signal-to-noise ratio $(s/n)_{\text{out-min}}$, as then $(s/n)_{\text{out-min}}$ can be correlated with an average observer's performance under prescribed average conditions.[9,10] Values of $(s/n)_{\text{out}}$ above the 100 percent articulation level[10] can in principle be determined with the aid of a sound-level meter. The same procedure holds if we are interested only in determining the presence or absence of a signal. Methods (c) and (d) above may be conveniently used, as they avoid the psychological difficulties inherent in aural reception, or method (b) may be tried, since it is the simplest situation where the ear is employed directly.

Not all the methods are equally effective or practical. Type (a), for example, requires a relatively broad band for intelligibility, while in (b) we must know the fundamental frequency or one of its harmonics—an indeterminate quantity in speech. The same is true of methods of type (c), except that now the signal is not without a definite fundamental, and may be a simple form of modulation—i.e., a sine wave. However, the fundamental frequency must be known for it to be possible spectrally to locate the narrow audio filter in the region of the desired component. Attention to the d.c. in methods (b) and (c) offers a way out of this difficulty at first glance, but these techniques suffer from other troubles, among them frequency drift, temperature changes, etc. Finally, method (d) may be safely used with either a wide or narrow filter but with a consequent loss of information as to wave form. In brief, the purpose of reception, i.e., either intelligence or detection of a signal, will largely dictate the choice of method, which in turn will be subject to the limitations mentioned above and to the further restriction that the sensitivity of all these procedures depends on the length of time available for the integration or "smoothing" of the data, a process affected electromechanically in (c) and (d) rather than by the ear-brain combination characterizing methods (a) and (b). We remark also that results for speech, which cannot be treated analytically, can be correlated experimentally with the experimental results for sinusoidal modulation, for which the theory has been developed.

We determine first the power in the signal components of the output of the discriminator (for f-m systems), or the second detector if a-m reception is used. Letting E_0 represent this output wave, we find the desired components from the Fourier development of $\langle E_0 \rangle$. Here the brackets $\langle\ \rangle$ indicate the statistical or ensemble average over the random variables describing the noise part of E_0.[11] For the periodic or signal part of the output we find then[12]

[8] J. H. Van Vleck and D. Middleton, "A theoretical comparison of the visual, aural, and meter reception of pulsed signals in the presence of noise," J. App. Phys. **17**, 940 (1946). Part I contains a discussion of these methods for the radar case; many of the remarks apply equally well here. In particular, see Sections I and IV of Part I.

[9] Examples of the ways this may be done in the visual (radar) case are discussed in some detail by J. W. Lawson and G. E. Uhlenbeck, *Threshold Signals*, to appear as Vol. **24** of the Radiation Laboratory Series.

[10] J. C. R. Licklider and S. J. Goffard, "Effects of impulsive interference upon AM voice communication," J. Acous. Soc. Am. **19**, 653 (1947). This paper contains further references to the acoustical aspects of the problem.

[11] Because the random process is assumed to be stationary—i.e., the mechanism causing the random variations of the current or voltage waves remains unchanged during the period (O, T) of observation, time averages (indicated by the brackets $\langle\ \rangle_{Av}$) and statistical averages (denoted by the $\langle\ \rangle$) are interchangeable in the limit $T \to \infty$. Furthermore, the average over a period T_0 of the modulation may be taken either before or after the ensemble average, since there is no correlation between signal and noise. See also Section 1 of reference 4, M. C. Wang and G. E. Uhlenbeck, "On the theory of the Brownian motion II," Rev. Mod. Phys. **17**, 323 (1945), and S. O. Rice, "Mathematical analysis of random noise," Bell Sys. Tech. J. **23**, 282 (1944).

[12] It should be pointed out that $\langle\langle E_0 \rangle\rangle_{Av}$ and $\langle\langle E_{0-\text{per}} \rangle\rangle_{Av}$ are not necessarily the same. In the case of amplitude modulation it is always possible to separate (at least analytically) the signal components following detection from the noise terms, al-

$$\langle E_0(t)_{\text{per}}\rangle = \sum_{k=-\infty}^{\infty} a_k \exp(ik\omega_a t) \equiv s_a(t),$$

$$\omega_a = 2\pi/T_0 \quad \text{and} \quad (2.1)$$

$$a_k = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} \exp(-ik\omega_a t)\langle E_0(t)_{\text{per}}\rangle dt,$$

and T_0 is the period of the modulation. The mean-square power in the kth component is

$$P_k = \langle s_a(t)^2\rangle_{Av} = \frac{\epsilon_k}{T_0^2}\left|\int_{-T_0/2}^{T_0/2}\langle E_0(t)_{\text{per}}\rangle \exp(ik\omega_a t)dt\right|^2,$$

$$\epsilon_0 = 1, \quad \epsilon_{k\geq 1} = 2, \quad (2.2)$$

and the mean total power is simply $\sum_k P_k$.

In a-m, or narrow band f-m, for which the entire low frequency output is passed, the signal power $P_s(=\sum_k P_k)$ is obtained from (2.2), and the mean power associated with the noise alone, including modulation components due to the signal but excluding the d.c. is

$$N_n^2 = \langle\langle E_0(t)^2\rangle - \langle E_0(t)\rangle^2\rangle_{Av}, \quad (2.3)$$

since $\langle E_0(t)^2\rangle$ represents the mean total power output. If we take P_s to be a measure of signal power when speech is sensed (a), or to be an indication of the presence or absence of a signal (methods $(b)-(d)$), we have for the output signal-to-noise ratio

$$(s/n)_I^2 = (P_s - P_{s-dc})/N_n^2 = \{\langle\langle E_0(t)_{\text{per}}\rangle^2\rangle_{Av} - \langle\langle E_0(t)_{\text{per}}\rangle\rangle_{Av}^2\}/\langle\langle E_0(t)^2\rangle - \langle E_0(t)\rangle^2\rangle_{Av}. \quad (2.4)$$

When the d.c. is included (2.4) is easily modified by adding $\langle\langle E_0(t)_{\text{per}}\rangle\rangle_{Av}^2$ to the numerator and $\langle\langle E_0(t)\rangle^2\rangle_{Av} - \langle\langle E_0(t)_{\text{per}}\rangle\rangle_{Av}^2$ to the denominator.

For *broad band* f-m, however, only part of the output noise accompanies the signal to be perceived, and so the mean noise power in this situation is

$$n_b^2 = \int_0^\infty W_0(f)\,|G_a(f)|^2 df, \quad (2.5)$$

where now $G_a(f)$ is the amplitude response of the audio filter and $W_0(f)$ is the power spectrum of the low frequency portion of the discriminator's output (or that from the second detector in a-m). The corresponding expression for the signal again depends on the mode of presentation. When the entire signal is observed, or detected by the combination bolometer and sound-level meter, for example, the total signal power (exclusive of d.c.) is chosen as a criterion [(a) and (d)], and we have here for the signal-to-noise ratio

$$(s/n)_{II}^2 = (P_s - P_{s-dc})/n_b^2 = \{\langle\langle E_0(t)_{\text{per}}\rangle^2\rangle_{Av} - \langle\langle E_0(t)_{\text{per}}\rangle\rangle_{Av}^2\} \Big/ \int_0^\infty W_0(f)\,|G_a(f)|^2 df. \quad (2.6)$$

If the d.c. is included, a modification similar to that for (2.4) may be made.

When only one component is heard (method (b)) the mathematical criterion for (s/n) is the ratio of signal power in this component to the mean noise power in the immediate spectral vicinity of this harmonic, where now the "immediate vicinity" is the width of the audio filter or the natural selectivity of the ear's filter response, whichever is narrower. (For further details see reference 8.) The harmonic usually selected is the first ($k=1$), and we have, accordingly, from (2.2) and (2.5)

$$(s/n)_{III}^2 = (2/T_0^2)\left|\int_{-T_0/2}^{T_0/2}\langle E_0(t)_{\text{per}}\rangle \exp(ik\omega_a t)dt\right|_{k=1}^2 \Big/ \int_0^\infty W_0(f)\,|G_a(f)|^2 df. \quad (2.7)$$

If, instead, the signal consists of but one component, and the entire low frequency is passed, our criterion (2.7) reduces to $(s/n)_I$, cf., Eq. (2.4). Meters (method (c)) are not considered here, but the remarks of Section IV, Part I (reference 8) and the analysis (Section V, Part II) are particularly pertinent and may be applied with little modification.

3. Representation of Signal and Noise in F-M Reception

The present theory is subject to a number of limitations, among them the requirement that the discriminator respond linearly for all frequencies. This restriction is not serious as long as the deviations of the carrier about resonance are not too great, or, if excessive, are of duration short relative to the time spent in the linear portion of the frequency-modulation voltage characteristic. Further, when there is limiting, the filter elements following the limiter must be sufficiently wide to pass the clipped wave without additional distortion due to frequency selection. Theoretically, an infinite pass band is assumed for the limiter; practically, a band width of the order twice the i-f is satisfactory,

though the mixture is no longer the simple superposition of signal and noise that entered the receiver. Thus we write symbolically $E_0(t) = E_0(t)_{\text{per}} + E_0(t)_{\text{noise}}$, and if $\langle E_0(t)_{\text{noise}}\rangle$ does not vanish (when the d.c. is filtered out), clearly $\langle E_0(t)\rangle_{Av} \neq \langle E_0(t)_{\text{per}}\rangle_{Av}$. On the other hand, as we shall see clearly in Section 4 (and Part II), it is possible that $\langle\langle E_0(t)\rangle\rangle_{Av} = \langle E_0(t)_{\text{per}}\rangle_{Av}$ in f-m reception, for then $E_0(t) = E_0(t)_{\text{per}} \cdot f(E_{0\,\text{noise}}; E_{0\,\text{carrier}})$, for example, Eq. (II.18) or (II.24): the output of the discriminator contains the modulation as a simple product with the noise distribution, which depends on the carrier but not on the modulation (for an ideal, distortionless discriminator). Similar remarks hold $\langle\langle E_0^2\rangle\rangle_{Av}$ *vs.* $\langle\langle E_0^2{}_{\text{per}}\rangle\rangle_{Av}$.

since even in cases of extreme limiting the amount of energy outside the original spectral width, due to clipping of the noise and signal, is small.[13] A third condition is that the noise be of the fluctuation type attributable to thermal or shot effects, or to any other stochastic process that is Gaussian.[14] Non-periodic interference, consisting of intermittent random pulses, i.e., "static," ignition noise, etc., cannot be treated in the same way mathematically, as their distribution laws are not yet known.

Because of the narrow band property of the signal and noise we may write for the wave leaving the i-f filter[15]

$$\begin{aligned} V(t) &= A_0\cos(\omega_c t+\Psi)+V_N(t) \\ &= [V_c+A_0\cos(\omega_d t+\Psi)]\cos\omega_0 t \\ &\quad -[V_s+A_0\sin(\omega_d t+\Psi)]\sin\omega_0 t, \end{aligned} \tag{3.1}$$

where A_0 is the peak carrier amplitude and $\omega_d/2\pi$ is the amount of detuning (c.p.s.), if any, between the carrier frequency (shifted to the i-f band) and the resonant frequency $f_0(=\omega_0/2\pi)$ of the i-f. The quantities V_c and V_s are the slowly varying components of the noise in phase with $\cos\omega_0 t$ and $\sin\omega_0 t$, respectively, and $\Psi(t)$ is the time integral of the modulation $D_0(t)$, *viz*:

$$\Psi(t)=\int^t D_0(t)dt. \tag{3.2}$$

The narrow band requirement means that Ψ and $\dot{\psi}(=D_0(t))$ do not contain significant spectral components outside the i-f band and that $\omega_d \ll \omega_0$ also. Further, it is necessary that the maximum deviation of the carrier be not too great, as otherwise amplitude modulation will occur, introduced by the excursion of the carrier into the low gain portions of the i-f filter response.

The disturbance $V(t)$ can be represented in the form of an envelope and phase:

$$V(t)=R(t)\cos[\omega_0 t+\theta(t)], \tag{3.3}$$

where R is the envelope and θ a phase, both of which are slowly varying functions of time compared to $\omega_0 t$. Comparison of (3.1) and (3.3) enables us to write at once

$$\begin{aligned} &R=[(V_c+\alpha)^2+(V_s+\beta)^2]^{\frac{1}{2}}; \\ &\theta=\tan^{-1}[(V_s+\beta)/(V_c+\alpha)]+m\pi, \\ &\qquad\qquad m=0,\ \pm1,\ \pm2,\ \cdots, \end{aligned} \tag{3.4}$$

in which

$$\alpha\equiv A_0\cos(\omega_d t+\Psi) \quad\text{and}\quad \beta\equiv A_0\sin(\omega_d t+\Psi). \tag{3.4a}$$

The low frequency output $E_0(t)$ of our ideal discriminator is given by Eq. (II.6). The effect of the limiter is contained in the envelope term $B_1(R)$, the precise value of which depends on our choice of dynamic characteristic $g(V)$ for the limiter. The quantity $f(iz)$, which is the Fourier transform of $g(V)$ as indicated by Eq. (II.1), is found to be for the linear limiter saturating at $|V|=R_0$:

$$f(iz)=2\eta(1-\exp(-iR_0 z))/(iz)^2. \tag{3.5}$$

The factor 2 appears because both positive and negative portions of the high frequency wave contribute to the envelope, while η is the dynamic transfer constant of the limiter. The envelope $B_1(R)$ is readily shown[16] to have the values

$$\left.\begin{aligned} B_1(R)_M &= \eta R, \quad 0\le R\le R_0. \\ &= \frac{4}{\pi}\eta R_0\,{}_2F_1(-1/2,\,1/2;\,3/2;\,R_0^2/R^2) \\ &= \frac{2\eta R}{\pi}\left[(R_0/R)\left(1-\frac{R_0^2}{R^2}\right)^{\frac{1}{2}}+\sin^{-1}(R_0/R)\right] \end{aligned}\right\} R_0\le R. \tag{3.6}$$

In extreme or "super"-limiting, for which R_0 is much less than the r.m.s. noise voltage b_0 leaving the i-f (i.e., $R_0^2 \ll 2b_0$), we retain only the first terms of (3.6), so that

$$[B_1(R)_M]_{\text{extreme}} \doteq 4\eta R_0/\pi, \quad (R_0^2\ll 2b_0). \tag{3.7}$$

Blachman's expressions for $B_1(R)$ are[17]

$$\left.\begin{aligned} B_1(R)_B &= \eta R, \quad 0\le R\le R_0 \\ &= \eta R_0, \quad R_0\le R \end{aligned}\right\}. \tag{3.8}$$

It is immediately evident that the two representations are identical when there is no limiting ($R_0\to\infty$) and differ by a factor $4/\pi$ (2.1 db) in extreme limiting. Thus, the approximate form $B_1(R)_B$ is always too small by an amount $\sigma(1\le\sigma\le 4/\pi)$.

4. Narrow Band F-M

To calculate signal-to-noise ratios at the discriminator output when narrow band f-m is employed we must determine the (statistical) average and mean-square values of $E_0(t)$ and apply the results to Eqs. (2.4), (2.6), or (2.7), as the situation

[13] D. Middleton, "The response of biased, saturated linear and quadratic rectifiers to random noise," J. App. Phys. **17**, 778 (1946). See Fig. 10.

[14] Reference 6, Section 1.

[15] Reference 4, Eqs. (2.3)–(2.10).

[16] The integration is accomplished by representing $\exp(-iR_0 z)$ as the sum of two Bessel functions and evaluating the resulting Weber-Schafheitlin integrals with the help of (A3.18) and (A3.19) of reference 6.

[17] Reference 3, Sections 2 and 3. The subscripts M and B are used to distinguish the accurate and approximate forms of the limited envelope, respectively.

warrants. From (II.6) the average is

$$\langle E_0(t)\rangle = \langle B_1(R)\dot{\theta}\rangle, \quad \dot{\theta}=\frac{d\theta}{dt},$$

$$= \kappa \int_0^\infty B_1(R)dR \int_{-\infty}^\infty \dot{\theta} W_1(R, \dot{\theta}; t)d\dot{\theta}, \quad (4.1)$$

where $W_1(R, \dot{\theta}; t)$ is the probability density for R and $\dot{\theta}$; (the time may appear explicitly because of the modulation). In a similar fashion we represent the mean-square output by

$$\langle E_0(t)^2\rangle = \kappa^2\langle B_1(R)^2\dot{\theta}^2\rangle$$

$$= \kappa^2 \int_0^\infty B_1(R)^2 dR \int_{-\infty}^\infty \dot{\theta} W_1(R, \dot{\theta}; t)d\dot{\theta}. \quad (4.2)$$

To obtain $W_1(R, \dot{\theta}; t)$ for (4.1) and (4.2) above we must first establish $W_1(R, \dot{R}, \theta, \dot{\theta}; t)$ for the envelope, phase, and their derivatives, of a modulated carrier and (Gaussian) random noise, since

$$W_1(R, \dot{\theta}; t) = \int_{-\infty}^\infty d\dot{R} \int_0^{2\pi} d\theta W_1(R, \dot{R}, \theta, \dot{\theta}; t). \quad (4.3)$$

The quantity $W_1(R, \dot{R}, \theta, \dot{\theta}; t)$ has already been determined[18] and is given in Part II, where the averages (4.1) and (4.2) are also evaluated. The final results are contained in normalized form in Part II along with the limiting forms of the results associated with various extreme values of the parameters. Numerical integration must be used in the general case, since it is not possible to obtain $\langle E_0\rangle$ or $\langle E_0^2\rangle$ in closed form except in the instances of no limiting ($r_0\to\infty$) or superlimiting ($r_0^2\ll 1$).

Figure 1 illustrates the mean low frequency output $\langle E_0\rangle$ as a function of the input carrier-to-noise ratio a_0 for a wide variety of limiting levels (r_0). As we expect, increased limiting results in a decreased output, since only a portion of the input wave is passed to enter the discriminator. Further, for a given amplitude of clipping, saturation occurs when the incoming carrier is sufficiently strong, as then there can be no further increase in the effective amplitude (determined by the limiter). For weak signals, however, we observe a rapid falling-off of the d.c. output, proportional now to the *square* of the input ratio a_0. This is the familiar phenomenon of *modulation suppression*, whereby the relatively stronger noise overwhelms the weaker carrier in non-linear fashion, i.e., proportional to a_0^2, instead of to a_0. A precisely similar effect is noted in a-m reception under the same circumstances of weak signal and strong noise.[7] See Eqs. (4.9) and (4.10) following. We observe that in all cases the approximate theory yields a smaller value of $\langle E_0\rangle$ than does the more precise analysis (by 2.1 db or less).

Next, from Fig. 2 we see how the mean-square noise output $\langle E_0^2\rangle$ depends on the carrier, when there is no modulation ($\phi=0$) (Eq. (II.9)). Suppression of the noise is now apparent as the carrier becomes more intense and the limiting is maintained constant. A similar effect is seen to occur when the degree of limiting is increased ($r_0\to 0$) and the incoming carrier a_0 is held fixed. Only when there is no limiting is there no absolute suppression of the noise; the mean low frequency output power remains unchanged as $a_0\to\infty$ (*vide* Fig. 4, reference 4).

It is convenient to introduce the quantity Γ, defined by

$$\Gamma \equiv \{\langle \dot{\phi}(t)^2\rangle_{Av}/\rho^2\}^{\frac{1}{2}}, \quad (4.4)$$

which is the ratio of the r.m.s. modulation to the *band width factor* ρ (defined in Eqs. (II.9) and (II.16)). As a satisfactory filter characteristic for the i-f stage we choose the Gaussian response throughout;[19] this means that $w(f)$ in (II.9) is

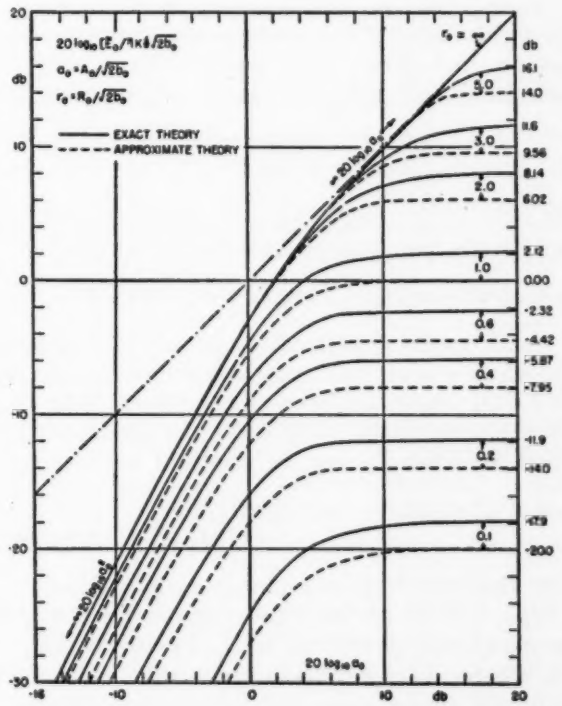


FIG. 1. Signal output of the discriminator (narrow band f-m) as a function of carrier strength for various levels of limiting.

[18] D. Middleton, *Spurious Signals Caused by Noise in Triggered Circuits*, Tech. Report No. 24, Dec. 10, 1947, Cruft Laboratory, J. App. Phys. 19, 817 (1948). See in particular Eqs. (II-5.4)–(II-5.7) of the Report, which apply here if β and ψ therein are changed to $-\beta$, $-\psi$

[19] The Gaussian response, while not physically realizable (H. Wallman, *Realizability of Filters* (M.I.T.), Radiation Laboratory Report No. 637 (Dec. 8, 1944)), is nevertheless a better model in most cases than the so-called "ideal" or

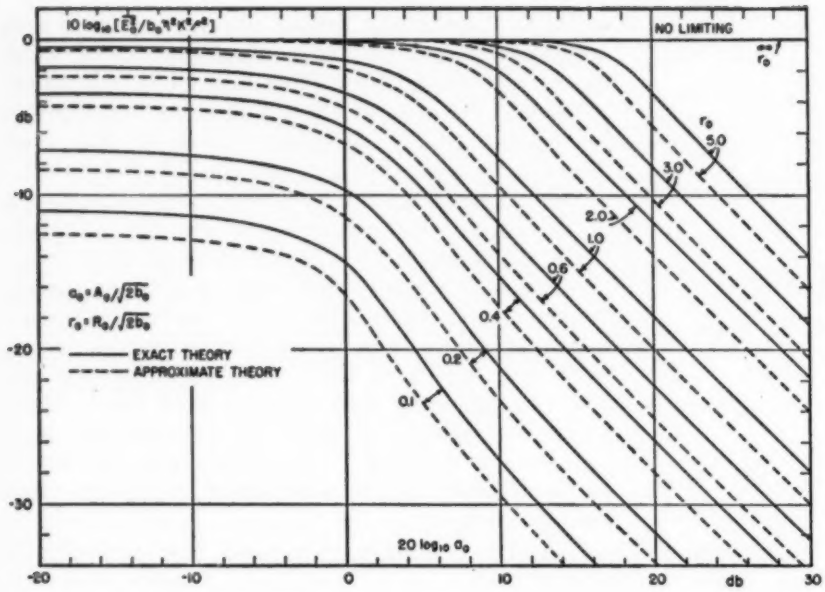


FIG. 2. Noise power output of the discriminator (narrow band f-m) as a function of carrier strength for various levels of limiting. No modulation ($\phi=0$).

given by

$$w(f)=W_0\exp(-\omega^2/\omega_b^2),\quad \omega=2\pi f,\quad \omega_b=2\pi f_b, \quad (4.5)$$

and here W_0 is the maximum spectral intensity (power/frequency) and $\omega_b(=2\pi f_b)$ is an angular frequency directly proportional to the width $(\Delta f)_{\frac{1}{2}}$, measured between half-power points. In terms of the width (Δf_E) of the equivalent rectangular filter passing the same mean power and having the same maximum spectral intensity W_0, we may write

$$f_b=\Delta f_E/(\pi)^{\frac{1}{2}};\quad (\Delta f)_{\frac{1}{2}}=((\ln 16)/\pi)^{\frac{1}{2}}\Delta f_E;$$
$$(\Delta f)_{\frac{1}{2}}=(\ln 16)^{\frac{1}{2}}\cdot f_b, \quad (4.6)$$

and from (4.5) and (II.9), with the definition of ρ, Eq. (II.16), we obtain for the Gaussian filter

$$\rho=\omega_b/\sqrt{2}=\pi\sqrt{2}f_b=(2\pi)^{\frac{1}{2}}\Delta f_E=\pi(\Delta f)_{\frac{1}{2}}/(\ln 4)^{\frac{1}{2}}. \quad (4.7)$$

If the frequency modulation consists of a single sine-wave whose maximum amplitude is D_0, *viz.*, $\phi=\psi=D_0\cos\omega_a t$, we have at once from (4.4)–(4.7) the following relations between the maximum (angular) deviation D_0 and the width of the i-f filter

$$D_0=\rho\Gamma\sqrt{2}=2\pi f_b\Gamma=\Gamma\omega_b=2(\pi)^{\frac{1}{2}}\Gamma\Delta f_E$$
$$=\pi\Gamma(\Delta f)_{\frac{1}{2}}/(\ln 2)^{\frac{1}{2}}. \quad (4.8)$$

Note that the *frequency* deviation is $D_0/2\pi$; thus, in Figs. 3–5, which have been drawn for $\Gamma=1/\sqrt{2}$, the maximum deviation is $D_0/2\pi=0.424(\Delta f)_{\frac{1}{2}}$ or $=0.399\Delta f_E$. This is not so great that appreciable amplitude modulation and consequent distortion of the signal, due to the traversal of the i-f characteristic by the carrier, is introduced.

When the input carrier is weak relative to the noise, the mean total output power is essentially independent of the carrier (Fig. 3), since the carrier is then suppressed by the noise. The output accordingly depends on the degree of limiting, being less for heavy clipping because less of the original wave is passed to the discriminator. As the carrier is increased, however, the contribution due to the modulation becomes significant and approaches a limit proportional to r_0, determined by the limiter. The purely noise contribution is suppressed, as (II.23) or (II.31) shows.

There remain the signal-to-noise ratios for narrow band f-m. These are shown in Fig. 4 and in more detail in Fig. 5, which corresponds to the region of the inset of the preceding figure. A number of interesting conclusions can be drawn from the figures immediately: first, limiting always decreases the sensitivity—i.e., raises the amount of input carrier power necessary to obtain a given value of the output ratio $(s/n)_I$. This is particularly evident in the threshold cases $(s/n)_I\leq 3$ db. Also, for strong carriers ($a_0^2>10$ db) the output $(s/n)_I$ is independent of the limiting, provided the clipping level is less than the peak carrier amplitude. The results of Blachman's approximate theory are only slightly smaller ($\frac{1}{2}$ db or less) in the ratio $(s/n)_I$ than the expressions of the more refined theory, in contrast with the considerably larger differences (up to 2.1 db) in the respective d.c., mean total power outputs, etc., when determined separately.

To compare $(s/n)_I$ with the analogous situation in a-m reception we need an expression for the signal-to-noise ratio $(s/n)_I$ corresponding to that

rectangular filter—itself unrealizable. In practice, cascaded, single-tuned circuits, for example, yield an over-all frequency response which approaches the Gaussian, and in f-m or a-m receivers there are enough such circuits to make our analytic choice a good one, even at large departures from resonance. Added to this is the greater mathematical convenience that Gaussian characteristics confer in the calculation of spectra.

defined in the f-m case. We obtain, finally,[20]

$$(s/n)_{I-\text{quad}}{}^2=a_0{}^4\lambda^2(2+\lambda^2/8)/[1+a_0{}^2(2+\lambda^2)], \quad |\lambda|\leq 1, \quad a_0\geq 0, \quad (4.9)$$

which applies for a sinusoidally-modulated carrier detected by a half-wave, unsaturated quadratic rectifier; λ is the modulation index. A similar relation, too involved to be given here, has been found[20] for the half-wave linear rectifier. Figures 4 and 5 show the dependence of $(s/n)_{I-\text{lin}}$ and $(s/n)_{I-\text{quad}}$ on a_0 for 100 percent modulation ($\lambda=1$). When the carrier is sufficiently weak or strong relative to the noise, the ratios in the a-m case are easily written:

$$(s/n)_{I-\text{lin}}\doteq a_0{}^4\lambda^2(2+\lambda^2/8)/(16/\pi-4+(4/\pi-1/2)(2+\lambda^2)a_0{}^2) =a_0{}^4\lambda^2(2+\lambda^2/8)/(1.093+0.512a_0{}^2(2+\lambda^2)), \quad a_0{}^2<1/2, \quad (4.10)$$

and

$$(s/n)_{I-\text{lin}}{}^2\simeq a_0{}^2\lambda^2;$$
$$(s/n)_{I-\text{quad}}{}^2\simeq a_0{}^2\lambda^2(2+\lambda^2/8)/(2+\lambda^2), \; a_0{}^2>10. \quad (4.11)$$

Compare these expressions with the general results for the f-m receiver, where now Γ (Eq. (4.4)) may take on any (positive) value consistent with the requirement that no significant amplitude modulation be introduced, i.e., $\Gamma^2\leq 1/2$. Now when there is little or no limiting ($r_0\to\infty$) we can safely choose $\Gamma\leq 1/\sqrt{2}$, with an ignorable amount of a-m in most instances ($\tanh\Gamma^2=0.25=25$ percent). From Part II we find that expressions for $(s/n)_I$ corresponding to (4.10) and (4.11) are, for no limiting

$$(s/n)_I{}^2\doteq\pi a_0{}^4\Gamma^2/2(1+a_0{}^2\Gamma^2), \quad a_0{}^2<\tfrac{1}{2},$$

and

$$(s/n)_I{}^2\simeq 2a_0{}^2\Gamma^2, \quad a_0{}^2>10, \quad \Gamma^2\leq\tfrac{1}{2}, \quad (4.12)$$

and for heavy limiting ($r_0{}^2\ll 1$ or $r_0{}^2\ll a_0{}^2$) this ratio reduces to

$$(s/n)_I{}^2\doteq 2\Gamma^2a_0{}^4/(\Lambda(1+a_0{}^2\Gamma^2)+a_0{}^2(1-\Lambda^2)), \; a_0{}^2<1/2,$$

and

$$(s/n)_I{}^2\simeq 2a_0{}^2\Gamma^2/(1+\Gamma^2), \; a_0{}^2>10, \; \Gamma^2\leq 1/2, \quad (4.13)$$

where

$$\Lambda=\Lambda_M=0.2731-Ei(-r_0{}^2),$$

Eq. (II.27), or

$$\Lambda=\Lambda_B=1-Ei(-r_0{}^2). \quad (4.14)$$

A comparison between a-m and f-m is also illustrated in Figs. 4 and 5 for $\lambda=1$ and $\Gamma^2=1/2$. Note that for narrow band f-m this comparison is made under the same conditions of operation—the i-f and audio bands are identical, and only the

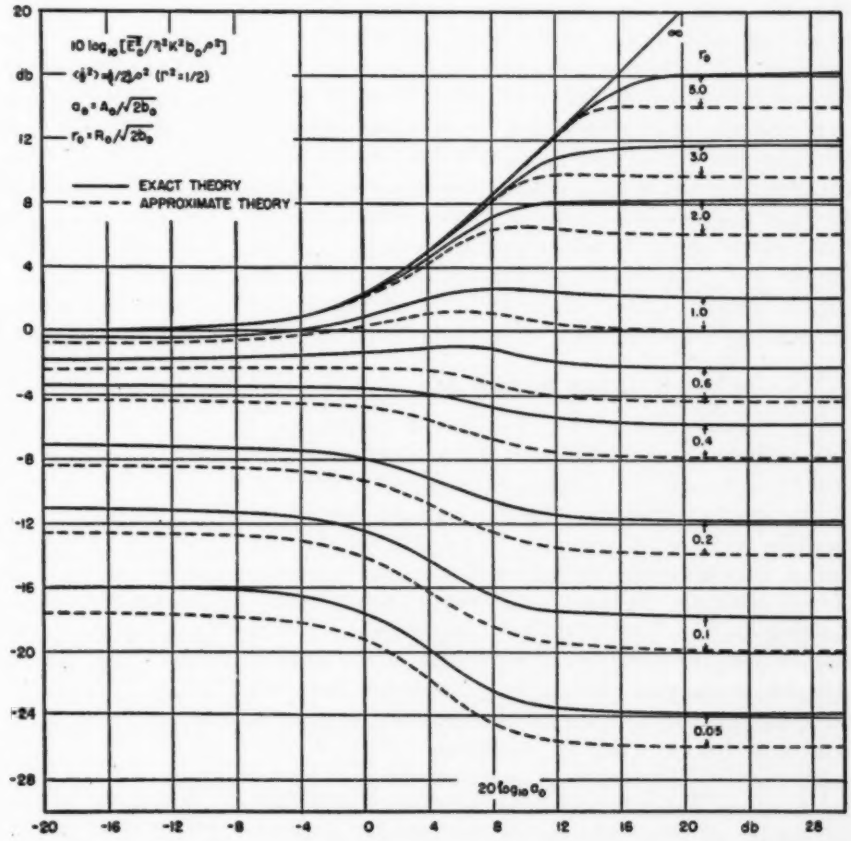


FIG. 3. Mean power output of the discriminator (narrow band f-m) as a function of carrier strength for various levels of limiting, with modulation:
$\langle\dot{\phi}^2\rangle_{Av}=\rho^2/2$. $\Gamma^2=1/2$.

[20] Unpublished work.

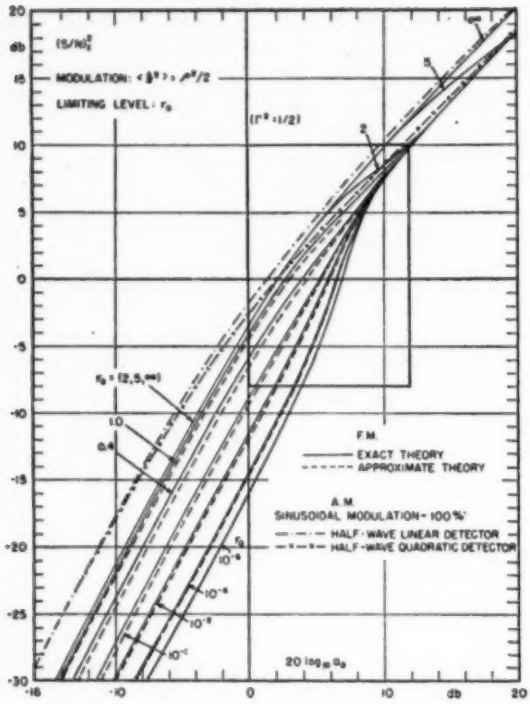


FIG. 4. Dependence of output signal-to-noise ratios (narrow band f-m) on input carrier level.

demodulating elements are different. We see, then, that for weak carriers f-m at its best (no limiting) is less effective than a-m by 3 db, while for strong carriers and no limiting the f-m receiver approaches the performance of the a-m receiver with linear detector. On the other hand, when there is heavy limiting the f-m device approaches in effectiveness the a-m receiver with a square-law detector. Intermediate degrees of limiting find the behavior of the f-m receiver somewhere between these two extremes, unless the clipping is especially heavy and the carrier weak. Then $(s/n)_I$ for f-m is noticeably diminished (*vide* Fig. 4).

5. Broad Band F-M

We consider next the problem of broad band f-m, which is characterized by a *deviation ratio* $\mu(\equiv D_0/\omega_a)$ usually much larger than unity. Because of the great complexity of the analysis for intermediate ranges of clipping $(0.1<r_0<\infty)$ we are forced to confine our attention to the two extremes of no limiting $(r_0\to\infty)$ and heavy or "super"-limiting $(r_0^2\ll a_0^2$, or $r_0^2\leq 10^{-2})$. But this is not too serious a restriction, since the values of the output signal-to-noise ratio follow monotonically between these limits, and so the two states together give us sufficient information to compare f-m and a-m performance. Nevertheless, our present problem is much more difficult than the narrow band case, since now we must determine the *spectral distribution* of the output noise and from this the portion of the total noise passed by the audio filter. A further complication arises because it is excessively tedious computationally to do this when there is modulation, except when the carrier is very strong[3,4] $(a_0^2>10$ db). The difficulty is ignorable, however, if the carrier is weak $(a_0^2<-3$ db), as then the results for the unmodulated case (see Section 3, reference 4) may be used, the effects of the modulation introducing only a comparatively small perturbation of the low frequency noise spectrum. The awkward region $(-3$ db$<a_0^2<10$ db) can accordingly be bridged by interpolation, with little error in most cases.

Since but a portion of the spectrum is passed by the final audio filter, the mean noise power n_b^2 accompanying the signal, all of whose components are assumed to lie in the audio pass-band, is given by Eq. (2.5). For convenience we consider the audio filter to be rectangular,[21] extending from 0 to f_A cycles $(f_A>f_a)$. Figures 6 (no limiting) and 7 (heavy limiting) show the integrated spectrum [based on Figs. 4 and 5 of reference 4 (paper)] or mean audio noise power, n_b^2, for various values of the upper cut-off frequency f_A. A significant feature of Fig. 6 is that n_b^2 approaches a constant value which is the same for *all* carrier strengths, when the audio filter becomes very wide, say $\Omega_A(\equiv f_A/f_b)\to 3$. This agrees with our result for narrow band f-m, which requires that the total output noise power (in the low frequency output of the discriminator) is independent of the carrier. On the other hand, when the clipping is heavy $(a_0^2\gg r_0^2;\ 10^{-2}\geq r_0^2)$ we see from Fig. 7 that the audio noise power n_b^2 still approaches a limit, now, however, dependent on the carrier intensity and on the degree of limiting; n_b^2 decreases proportional to a_0^{-2} $(a_0^2>10$ db) because of absolute suppression of the noise by the carrier.

The signal-to-noise ratio applicable for broad band f-m when all components except the d.c. are passed is given by Eq. (2.6), where $\langle E_0\rangle$ is obtained from (II.18). The mean noise power for no limiting is

$$n_b^2=\int_0^\infty W_0(f)_{r_0\to\infty}\cdot|G_a(f)|^2df$$
$$=\kappa^2\eta^2 b_0\omega_b^2\pi^{-\frac{1}{2}}N_b^2(\Omega_A;\infty)_1,\quad \Omega_A=f_A/f_b,\quad (5.1)$$

where $N_b^2(\Omega_A;\infty)_1$ is a (dimensionless) number obtained by integrating the spectra (for no limiting) between the limits 0 and f_A for the particular audio filter $G_a(f)$ considered above. We have, finally, for the desired ratio

$$(s/n)_{II}^2=\Gamma^2\pi^{\frac{1}{2}}a_0^4\exp(-a_0^2)[I_0(a_0^2/2)$$
$$+I_1(a_0^2/2)]^2/4N_b^2(\Omega_A;\infty)_1,\quad \Gamma^2\leq 1/2.\quad (5.2)$$

[21] Actually the response is more nearly Gaussian, but the error arising from the simplification is trivial if we replace the true filter by the equivalent rectangular one, according to Eq. (4.6).

As before for narrow band f-m the quantity Γ is required to be equal to or less than $1/\sqrt{2}$, in order to avoid serious amplitude modulation, and for optimum performance Γ is chosen as large as possible consistent with this requirement. It is possible to make Γ even greater (than $1/\sqrt{2}$) if we are willing to limit heavily. Unfortunately, most of the carrier energy is then wasted, falling as it does outside the normal acceptance band of the i-f. Furthermore, a large amount of noise is mixed with the modulation, due to the non-linear slope of the i-f characteristic, with the result that such a wave, although limited, is extremely noisy. The net effect is a considerable reduction in range and intelligibility. Our ratio (5.2) applies only if $a_0^2 < -3$ db, as explained above; for values of a_0^2 exceeding 10 db we may use (5.2) provided we replace $(N_b^2)_1$ by

$$N_b^2(\Omega_A;\infty)_2 \simeq \sum_{n=0}^{\infty} \epsilon_n J_n(\mu)^2 \times \left\{ \frac{n-1}{2} \exp(-(n-1)^2\Omega_A^2) - \frac{n+1}{2} \exp(-(n+1)^2\Omega_A^2) + \frac{(\pi)^{\frac{1}{2}}}{4}\Theta\{(n+1)\Omega_A\} - \frac{(\pi)^{\frac{1}{2}}}{4}\Theta\{(n-1)\Omega_A\} \right\},$$

$$\epsilon_0 = 1, \quad \epsilon_{n\geq 1} = 2, \quad (5.3)$$

obtained for our Gaussian i-f on specializing to sinusoidal-frequency modulation (cf. Eq. (3.30) of reference 4, $\omega_d = 0$). Here Θ is the well-known error function, tabulated in Jahnke and Emde:[22]

$$\Theta = \frac{2}{(\pi)^{\frac{1}{2}}} \int_0^x \exp(-y^2)dy; \quad \Theta(x) = -\Theta(-x).$$

The deviation ratio μ is related to Γ and the modulating frequency f_a by

$$\mu = D_0/\omega_a = D_0/\omega_b\Omega_a = \Gamma/\Omega_a, \quad (5.4)$$

since $\Gamma = \omega_b^{-1}D_0$ for the Gaussian i-f (Eq. (4.8)). In what follows, whenever sinusoidal modulation occurs, we choose $f_a = f_A$, so that μ becomes Γ/Ω_A.

We observe that for large values of a_0 the ratio (5.2) reduces to

$$(s/n)_{II}^2 \simeq (\pi)^{\frac{1}{2}}\Gamma^2 a_0^2/N_b^2(\Omega_A;\infty)_2, \quad a_0^2 > 10 \text{ db}, \quad \Gamma^2 \leq 1/2, \quad (5.5)$$

and thus $(s/n)_{II}^2$ is proportional to a_0^2, since by (5.3) $(N_b^2)_2$ is independent of the carrier, a result in accord with the conclusions of our earlier study.[4] When the audio filter is made sufficiently wide to include the entire low frequency output of the discriminator ($\Omega_A \to 3$ or more) we see that $(N_b^2)_2 \to (\pi)^{\frac{1}{2}}/2$, so that our ratio (5.5) reduces to the expected expression for narrow band f-m, *viz.* (4.12). A corresponding simplification is noted for weak carriers:

$$(s/n)_{II}^2 \doteq \Gamma^2\pi^{\frac{1}{2}}a_0^4/4N_b^2(\Omega_A;\infty)_1, \quad a_0^2 < -3 \text{ db}, \quad \Gamma^2 \leq \tfrac{1}{2}, \quad (5.6)$$

where once again suppression of the signal by the noise is observed (see Section 4 and Part II).

The dependence of the signal-to-noise ratio on carrier strength for various audio filter widths is

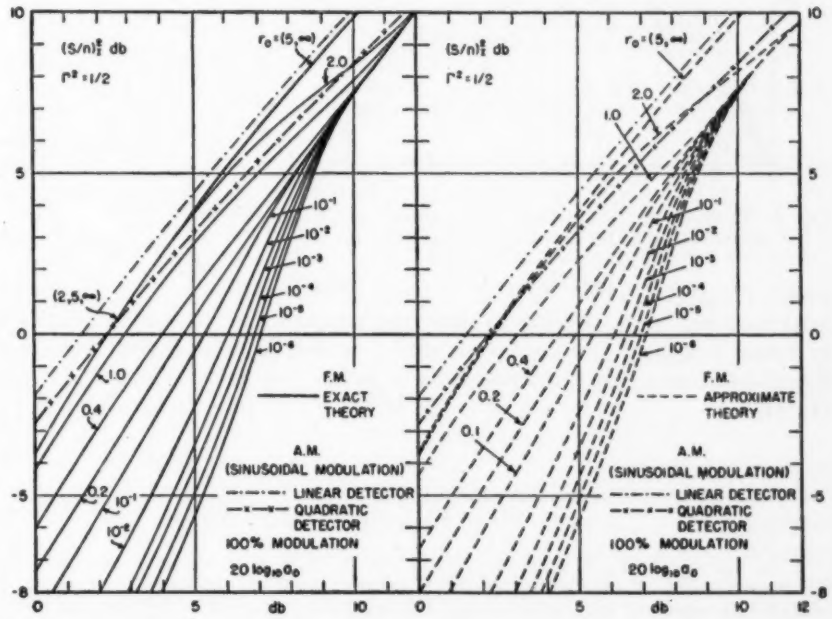


FIG. 5. Output signal-to-noise ratios (narrow band f-m).

[22] E. Jahnke and F. Emde, *Tables of Functions* (Dover Publications, New York, 1943).

illustrated in Fig. 8, where it is assumed that the i-f width is held fixed. Accordingly, narrowing the audio filter increases $(s/n)_{II}$, since less noise in the final stage accompanies the signal to be perceived. This assumes, of course, that all components of the signal still fall within the audio pass-band, so that as f_A is decreased, the spectral width of our signal is correspondingly modified, while the same mean carrier power is maintained at the input (i-f). However, it is even more important that we see what happens to $(s/n)_{II}$ when the audio filter width ($\sim\omega_A$), as well as the carrier power, is held constant, and various i-f widths ($\sim\omega_b$) are selected. Figure 9 illustrates such a comparison.[23]

Next, we examine the case of heavy limiting. The mean noise power (exclusive of d.c.) is here

$$n_b^2=\int_0^\infty W_0(f)_{r_0}\cdot|G_a(f)|^2df$$
$$=(16\kappa^2\eta^2/\pi^2)((\omega_b^2r_0^2b_0/(\pi)^{\frac{1}{2}}))N_b^2(\Omega_A;r_0)_1, \quad (5.7)$$

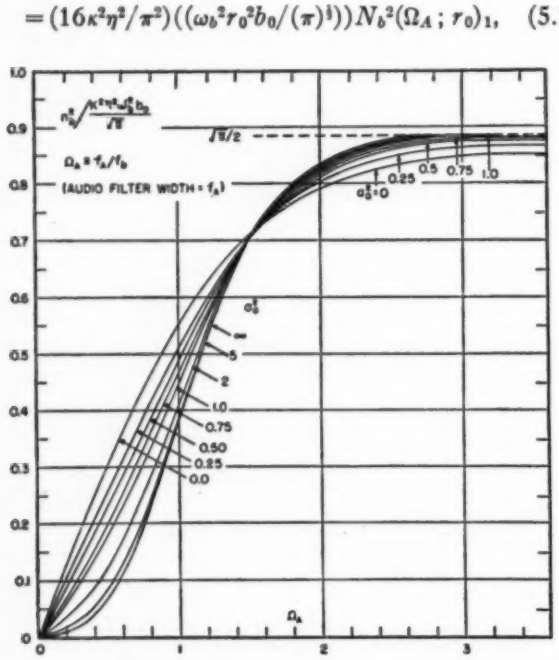


FIG. 6. Mean audio noise power from the discriminator after audio filtering (no limiting, and no modulation).

where $N_b^2(\Omega_A;r_0)_1$ like $N_b^2(\Omega_A;\infty)_1$ is a (dimensionless) number found by integrating the spectra (for extreme limiting) between 0 and f_A cycles. The signal-to-noise ratio (2.6) becomes with the aid of (II.24)

$$(s/n)_{II}^2=\Gamma^2(\pi)^{\frac{1}{2}}(1-\exp(-a_0^2))^2/N_b^2(\Omega_A;r_0)_1$$
$$\Gamma^2\leq 1/2;\quad (r_0\ll a_0^2;\ r_0^2\leq 10^{-2}). \quad (5.8)$$

For sufficiently heavy limiting this ratio is essentially independent of the clipping level, unless Ω_A becomes very large, in which case broad band f-m goes over into narrow band operation; even then the effect of limiting is not very marked, as the curves of Fig. 4 show. Figure 8 also illustrates $(s/n)_{II}$ as a function of a_0 for a constant i-f and variable audio band width; again (cf. Fig. 7) narrowing the audio decreases the amount of noise passed with the signal in the final stage, so that $(s/n)_{II}$ is increased. When the i-f is altered and the audio remains fixed, we obtain the curves shown in Fig. 9.

For carriers strong relative to the noise we must modify $N_b^2(\Omega_A;r_0)_1$; from Eq. (3.31) of reference 4 we find that for sinusoidal modulation

$$N_b^2(\Omega_A;r_0)_2\equiv a_0^{-2}\sum=a_0^{-2}\sum_{n=0}^{\infty}\epsilon_nJ_n(\mu)^2$$
$$\times\left\{\frac{n-1}{2}\Omega_A\exp(-(n+1)^2\Omega_A^2)\right.$$
$$-\frac{n+1}{2}\Omega_A\exp(-(n-1)^2\Omega_A^2)$$
$$+\frac{(\pi)^{\frac{1}{2}}}{2}[n^2\Omega_A^2+\tfrac{1}{2}][\Theta\{(n+1)\Omega_A\}$$
$$\left.-\Theta\{(n-1)\Omega_A\}]\right\},\quad a_0^2\gg 1, \quad (5.9)$$

so that (5.8) is

$$(s/n)_{II}^2\simeq(\pi)^{\frac{1}{2}}\Gamma^2a_0^2/\sum,\quad (a_0^2\gg 1),\quad \Gamma^2\leq 1/2, \quad (5.10)$$

which except for a numerical factor exhibits the same dependence on a_0 and Γ as does (5.5) when there is no limiting. The analog of (5.6) for weak carriers is found at once to be

$$(s/n)_{II}^2\doteq(\pi)^{\frac{1}{2}}\Gamma^2a_0^4/N_b^2(\Omega_A;r_0)_1,$$
$$a_0^2<-3\text{ db},\quad \Gamma^2\leq 1/2, \quad (5.11)$$

where similar suppression effects are observed.

After appropriate modification[20] of the results of reference 7 the signal-to-noise ratio (2.6) for a-m is found to be

$$(s/n)_{II-\text{quad}}^2=\frac{a_0^4\lambda^2(2+\lambda^2/8)}{\Theta(\Omega_a/\sqrt{2})+a_0^2[2\Theta(\Omega_a)+\lambda^2\Theta(\Omega_A+\Omega_a)/2+\lambda^2\Theta(\Omega_A-\Omega_a)/2]},\quad |\lambda|\leq 1 \quad (5.12)$$

[23] The curves of Fig. 8 may be used to obtain $(s/n)_{II}$, if we note that since $a_0\sim 1/(\omega_b)^{\frac{1}{2}}$ and $\Omega_A\sim\omega_b^{-1}$ we must choose the proper value of a_0 and Ω_A for the particular curve in question. Any curve (Ω_A=constant) may be used as reference; here, for example, we have chosen $\Omega_A=2.0$, which in terms of a rectangular audio and i-f means that $\Delta f_{\text{audio}}=2\Delta f_E/(\pi)^{\frac{1}{2}}=1.13\Delta f_E$, cf. Eq. (4.6). Thus, for an i-f 10 times as wide as our standard, we shift the curve in Fig. 8 for $\Omega_A=0.2(=2/10)$ to the right 10 db, etc.; a_0 is then still the independent variable.

which is applicable to half-wave quadratic rectifiers when a sinusoidally-modulated carrier is detected in random noise. Note that if the final audio filter is made large enough to include the entire low frequency spectrum ($\Theta \to 1$) Eq. (5.12) reduces to (4.9). The more complex formulas for the linear detector are not included here,[20] but curves showing $(s/n)_{II-\text{lin}}^2$ are given in Figs. 8 and 9. For small signals $(s/n)_{II-\text{quad}} \doteq (s/n)_{II-\text{lin}}$ quite closely (within a few tenths of a decibel), while for large carriers we obtain

$$(s/n)_{II-\text{quad}}^2 \simeq a_0^2\lambda^2(2+\lambda^2/8)/[2\Theta(\Omega_A) + \lambda^2\Theta(\Omega_A+\Omega_a)/2+\lambda^2\Theta(\Omega_A-\Omega_a)/2];$$
$$(s/n)_{II-\text{lin}}^2 \simeq a_0^2\lambda^2/\Theta(\Omega_A). \quad (5.13)$$

A similar procedure may be followed for $(s/n)_{III}$, cf. (2.7), where only one component of the signal is observed. Here it is the *spectral intensity* of the noise as a function of frequency that is important, since the width of the audio filter Δf_{audio} is very narrow compared to the frequency (kf_a) at which the desired component is located. As above, for weak carriers we may obtain the mean noise power n_b^2 accompanying the signal from the spectral distribution for unmodulated carriers and noise, *viz.*, Figs. 4 and 5 of reference 4:

$$\text{(no limiting)}\quad n_b^2 = \left(\kappa^2\frac{\eta^2\omega_b^2 b_0}{(\pi)^{\frac{1}{2}}}\right)\Delta\Omega_A \cdot W(\Omega_a;\infty)_1;$$

(extreme limiting)

$$n_b^2 = \left(\frac{16\kappa^2\eta^2 b_0 r_0^2\omega_b^2}{\pi^2(\pi)^{\frac{1}{2}}}\right)\Delta\Omega_A \cdot W(\Omega_a; r_0)_1$$

$$\Delta\Omega_A \equiv \Delta f_{\text{audio}}/f_b; \quad \Omega_a \equiv f_a/f_b, \quad (5.14)$$

where $W(\ )$ is the numerical value of the spectral intensity (seconds).

As a specific example we consider sinusoidal modulation; analogous to (5.2) the appropriate signal-to-noise ratio is

$$(s/n)_{III}^2 = \Gamma^2\pi^{\frac{1}{2}}a_0^4\exp(-a_0^2)[I_0(a_0^2/2) + I_1(a_0^2/2)]^2/4W(\Omega_a;\infty)_1\Delta\Omega_A, \quad \Gamma^2 \le 1/2, \quad (5.15)$$

while for heavy limiting, cf. (5.8), we have

$$(s/n)_{III}^2 = \Gamma^2(\pi)^{\frac{1}{2}}(1-\exp(-a_0^2))^2/W(\Omega_a; r_0)_1\Delta\Omega_A, \quad (r_0^2 \le 10^{-2} \text{ or } r_0 \ll a_0^2), \quad \Gamma^2 \le 1/2. \quad (5.16)$$

For strong carriers we replace $W(\Omega_a;\infty)_1$ and $W(\Omega_a; r_0)_1$, respectively, by

$$W(\Omega_a;\infty)_2 = \Omega_a^2\sum_{n=0}^{\infty}\epsilon_n J_n(\mu)^2 \times[(n+1)^2\exp(-(n+1)^2\Omega_a^2) + (n-1)^2\exp(-(n-1)^2\Omega_a^2)], \quad a_0^2 \gg 1 \quad (5.17)$$

and

$$W(\Omega_a; r_0)_2 = a_0^{-2}\Omega_a^2\sum_{n=0}^{\infty}\epsilon_n J_n(\mu)^2 \times[\exp(-(n+1)^2\Omega_a^2)+\exp(-(n-1)^2\Omega_a^2)], \quad (r_0^2 \le 10^{-2} \text{ or } r_0^2 \ll a_0^2). \quad (5.18)$$

These expressions follow from Eqs. (3.30) and (3.31)

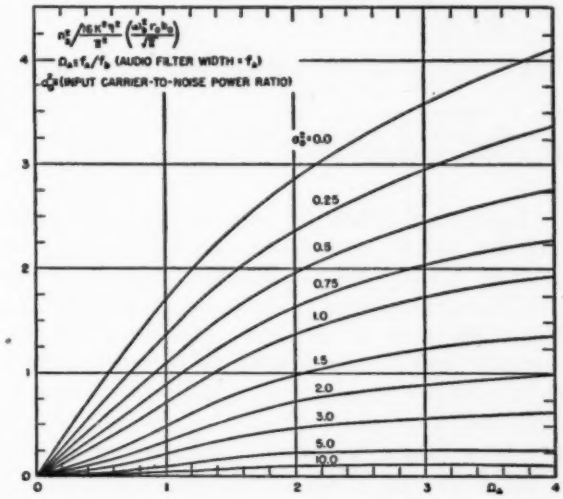


FIG. 7. Mean audio noise power from the discriminator after audio filtering (extreme limiting and no modulation).

of reference 4. The limiting behavior for small and large carriers is precisely the same as for the results (5.5), (5.6), (5.10), and (5.11), save that N_b^2 is replaced by $W(\ )\Delta\Omega_A$. The corresponding signal-to-noise ratios for a-m are obtained from

$$(s/n)_{III-\text{quad}}^2 = \frac{(\pi)^{\frac{1}{2}}a_0^4\lambda^2(2+\lambda^2/8)}{2\Delta\Omega_A[2^{-\frac{1}{2}}\exp(-\Omega_a^2/2)+2a_0^2\exp(-\Omega_a^2)+a_0^2\lambda^2(\exp(-4\Omega_a^2)+1)/2]}, \quad |\lambda| \le 1 \quad (5.19)$$

which is valid for all a_0; the ratio for the linear detector is essentially given by (5.19) when $a_0^2 \ll 1$ and for large carriers is

$$(s/n)_{III-\text{lin}} \simeq (\pi)^{\frac{1}{2}}a_0^2\exp(\Omega_a^2))/2\Delta\Omega_A, \quad (a_0^2 \gg 1). \quad (5.20)$$

We observe at once from Fig. 8 that under conditions of constant i-f and variable audio band widths (passing always the same signal power) a-m with a linear detector is superior in the threshold and small-signal regions ($(s/n)_{II} \lesssim 3$ db) to broad band f-m with or without limiting. However, when the incoming carrier is sufficiently intense ($a_0 > 3$ db) f-m begins to be more effective than a-m. In this

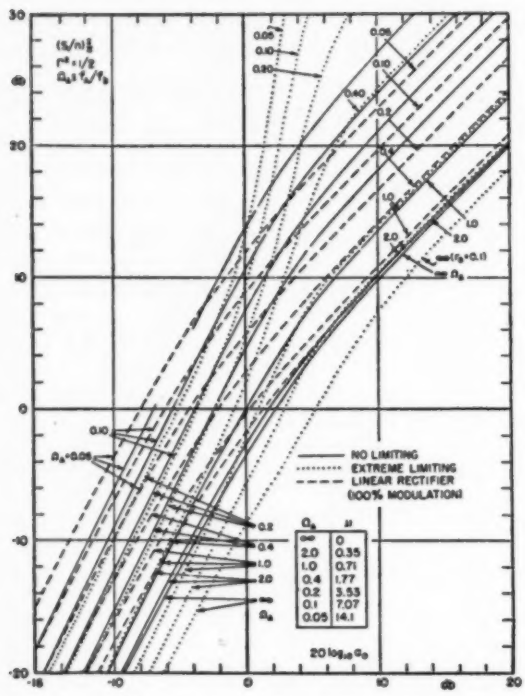


FIG. 8. Output signal-to-noise ratios for broad band f-m after audio filtering (sinusoidal modulation, constant i-f, and variable audio widths).

state f-m without limiting appears to be about 4 db better, for narrow audio filters $\Omega_A \simeq 0.4$ and $a_0^2 > 10$ db, while with heavy limiting there is a very marked improvement over a-m, and over f-m without limiting, under these conditions. The improvement occurs because the noise power accompanying the strong signal is greatly diminished in the narrow audio band $(0, f_A)$ due to the action of limiting, which redistributes the greater part of the noise energy into components beyond the upper cut-off of the audio filter. The noise spectrum in the band $(0, f_A)$ is nearly all attributable to $(s \times n)$ noise products[4] and is proportional to f^2, with a vanishing intensity at $f=0$. On the other hand, when there is no limiting the spectrum does not in general exhibit this diminution at the low frequencies, and the reduction in the noise intensity arises purely from suppression by the carrier, a suppression equally effective in a-m reception[7] when the carrier is strong. See Section 2, reference 4, cf. Eq. (2.42) and (2.47). As the audio filter is widened we lose this advantage of spectral shape, inherent in broad band f-m *with* limiting, and approach the condition of narrow band operation, for which limiting is actually detrimental (see Section 4). Similar remarks apply to $(s/n)_{III}$.

Comparison of f-m with a-m should normally be made under the optimum conditions of operation of the respective systems on the basis of the same standard of performance. This means in the broad band cases much wider i-f bands than would be used in a-m reception. For a-m we require an i-f sufficient to pass the highest component of the modulation with little attenuation; in broad band f-m, on the other hand, the deviation of the carrier from resonance is much greater than the conventional i-f width in amplitude-modulation systems, so that a correspondingly broad i-f is needed, as otherwise many important signal components would be lost. Figure 9 has been drawn with the above in mind and shows the output signal-to-noise ratios obtainable for the same modulating band widths and modulation power, with various i-f bands. As our standard of comparison[23] we have chosen the curve (in Fig. 8) for which $\Omega_A = 2.0$, corresponding to an audio width approximately equal to that of the i-f. In f-m this is essentially narrow band operation (Figs. 4 and 5), while in a-m it is the normal and optimum condition for communication purposes. We see at once that for thresholds of +6 db or less, a-m with a half-wave linear rectifier is superior to other types, requiring less carrier power than either broad or narrow band f-m, with or without limiting. Furthermore, even when the carrier is strong, a-m holds its own, except in the cases for which the i-f is made very wide (f-m) and extreme limiting is employed. In these instances it is the effect of limiting (as explained in the preceding paragraph) which gives heavily clipped, broad band f-m its advantage over the other systems. Note, however, that these effects become significant only at input carrier levels that are already high (>12 db) compared with the incoming noise, so that they contribute little towards a gain in *range*, for which the threshold, say $(s/n)_{II} \sim 0$ db, may require but a slight amount more incoming signal than noise, i.e., $(a_0^2 \sim 3$ db), *vide* Fig. 9.

6. Concluding Remarks

From Sections 4 and 5, in which the detailed explanations are given, we may make a few pertinent observations concerning the relative merits of narrow band f-m, broad band f-m, and a-m, from the point of view of reception of signals in fluctuation noise. In brief, narrow band f-m is inferior to broad band f-m for all input carrier intensities. Furthermore, limiting in the former is detrimental; the best conditions of operation require no limiting at all, but since in general practice the receiver must also operate against impulse noise, some limiting must be introduced, with a comparative reduction in optimum behavior (see Figs. 4 and 5 and compare with Fig. 9). Broad band f-m with heavy limiting provides the best performance, provided the incoming carrier-to-noise ratio is sufficiently large (*vide* Figs. 8 and 9), but is inferior to the case of no clipping when the carrier is weak. These results are

valid in the more usual situation for which the audio band width is fixed, and various i-f widths are chosen to fit the respective optimum conditions of operation. (Fig. 9). It should be remembered, of course, that these remarks apply for ideal discriminators (Section 3) and sufficiently wide limiter responses; departure from these conditions will ordinarily operate to increase the interference effects of the noise *versus* the signal, since filter characteristics narrow compared to the spectrum of the amplitude-limited noise tend to restore randomness[24,25] to the clipped wave and hence yield larger noise peaks, destroying the limiter's action. Inasmuch as the superiority of broad band f-m depends heavily on limiting, care should be taken to provide limiter and discriminator characteristics[26] wide compared to the i-f. In addition to the above, the carrier deviation should be as large as possible consistent with the requirement that ignorable amplitude modulation be introduced by the carrier sweeping through the i-f; if this is not so, the i-f then behaves from the viewpoint of the succeeding stages somewhat like a very non-linear discriminator, with a resultant serious mixing of the noise and signal.

Comparison of f-m and a-m according to the various conditions specified in Sections 4 and 5 indicates that *amplitude-modulation is more effective than either type of f-m for the reception of weak or even moderate signals in fluctuation noise*[27] (cf. Figs. 8–9). Only for strong signals and heavy limiting does broad band f-m give a noticeable improvement over a-m. Thus, from the point of view of range, which is set by the conditions of observation and usually requires an output signal-to-noise ratio no greater than a few decibels, a-m is to be preferred. a-m and narrow band f-m without limiting compare quite closely for strong signals (within a few tenths db).[28] These remarks are based on the theory as developed for sinusoidal modulation. When the frequency modulation is speech or music similar conclusions are predicted, but with somewhat higher values of the threshold region, due to the stricter conditions implied in the greater fidelity demanded at the output. Some betterment of the performance of broad band f-m relative to a-m reception can be expected from the use of pre-emphasis and de-emphasis, but it is doubtful whether this would be sufficient to compete with a-m when the signal is weak.

II. MATHEMATICAL SECTION

If we let $f(iz)$ be the Fourier transform of the limiter's dynamic characteristic $g(V)$, i.e.,

$$f(iz)=\int_0^\infty \exp(-iVz)g(V)dV, \quad Im(z)<0, \quad \text{(II.1)}$$

we may write the output $V_0(t)$ of the limiter from Eq. (3.3), as[29] where

$$V_0(t)=(1/2\pi)\int_C f(iz)\exp(-izR\cos(\omega_0 t+\theta))dz$$

$$=\sum_{n=0}^{\infty} B_n(R)\cos n(\omega_0 t+\theta), \quad \text{(II.2)}$$

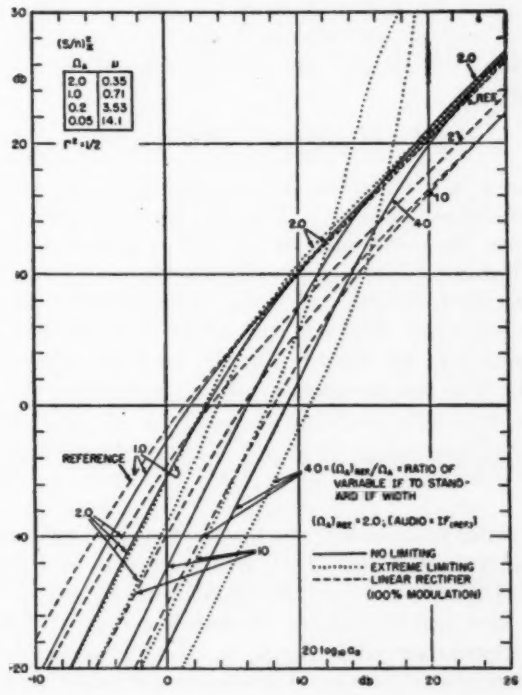


FIG. 9. Output signal-to-noise ratios for broad band f-m after audio filtering (sinusoidal modulation, variable i-f, and constant audio widths).

[24] S. O. Rice, "Mathematical analysis of random noise," Bell Sys. Tech. J. **24**, 46 (1945).

[25] M. Kac and A. J. F. Siegert, "On the theory of noise in radio receivers with square law detectors," J. App. Phys. **18**, 383 (1947).

[26] In practice the limiter is usually of low enough Q to satisfy this requirement easily; the discriminator response, however, is more critical but can be made linear for frequencies up to about twice the maximum carrier deviation, so that our rather idealized model is really quite satisfactory.

[27] Nothing has been said here about impulse noise, but from earlier work (see reference 5) it would appear that a-m reception with automatic volume control or otherwise properly regulated limiter gives as good performance under similar conditions as does f-m, and better, if the signal is weak. Further, f-m suffers far more from tuning difficulties and "capture-effects" than does a-m.

[28] Crosby (see reference 2) obtains a 4.8 db ($=\sqrt{3}$) superiority of narrow band f-m over a-m with a linear detector. The discrepancy between his and the present results is explained by the following two factors: (a) Crosby chooses a rectangular i-f pass-band, and we select a Gaussian one; this diminishes the effect of his noise *vs.* ours by $(3/2)^{\frac{1}{2}}$ or 1.8 db. (b) Also, with a rectangular i-f one can in theory tolerate a value of Γ equal to unity (Eq. 4.4) *vs.* the value $\Gamma=1/\sqrt{2}$ assumed for the Gaussian response; accordingly, for this larger maximum deviation he gets a further improvement of 3.0 db, making 4.8 db in all.

[29] See Eq. (1.8) of reference 6.

where

$$B_n(R)=(\epsilon_n i^n/2\pi)\int_C f(iz)J_n(Rz)dz,$$

$$\epsilon_0=1,\quad \epsilon_{n\geq 1}=2. \quad \text{(II.3)}$$

The contour **C** extends along the real axis from $-\infty$ to $+\infty$ and is indented downward about a possible singularity at $z=0$. By Cauchy's theorem this contour is equivalent to the (straight line) contour **C′** extending from $-\infty-i\alpha$ to $+\infty-i\alpha(\alpha>0)$ implied in the transform relation (II.1). The quantities $B_n(R)$ are the envelopes of the $n(=0,\ 1,\ 2\cdots)$ spectral zones that are produced in the limiter by its non-linear action;[6,13] $n\theta$ is the phase relative to $n\omega_0 t$, the nth harmonic of the central (i-f) frequency f_0. Only that portion of the wave after limiting which has the instantaneous phase $\omega_0 t+\theta$ is passed by the limiter's filter to enter the discriminator. This is the spectral band distributed about $f_0(n=1)$. If we represent the input vectorially by $\mathbf{V}=Ve^{i\Theta}$, for which $V=B_1(R)$ and $\Theta=\omega_0 t+\theta$, the time derivative of **V** in polar coordinates (with unit vectors **r**, **θ**) is

$$\dot{\mathbf{V}}=\mathbf{r}\dot{V}+\boldsymbol{\theta}V(d\Theta/dt);\quad (\mathbf{r}=e^{i\Theta};\quad \boldsymbol{\theta}=ie^{i\Theta}), \quad \text{(II.4)}$$

of which $V(d\Theta/dt)$ is clearly the time derivative of the phase or instantaneous frequency.[30] Now the output E_0' of our idealized discriminator (before filtering) is simply the coefficient of **θ** above, namely,

$$E_0(t)'=\kappa V(d\Theta/dt)=\kappa B_1(R)(\dot{\theta}+\omega_0), \quad \text{(II.5)}$$

where κ is a constant of proportionality with the dimensions (seconds). The i-f term, represented here as the coefficient of ω_0, is not passed by the discriminator filter, since the discriminator output is directly proportional to the instantaneous *difference* frequency between the wave and the central frequency f_0, of the i-f, limiter, and discriminator bands. The low frequency output is, finally,

$$E_0(t)=B_1(R)\dot{\theta}=(i\kappa\dot{\theta}/\pi)\int_C f(iz)J_1(Rz)dz. \quad \text{(II.6)}$$

In order to calculate the first and second moments of E_0, *viz.*, $\langle E_0\rangle$ and $\langle E_0^2\rangle$, which appear as fundamental quantities in the signal-to-noise ratio, we need first the first-order probability density function associated with R, $\dot{R}$, θ, and $\dot{\theta}$, when there is a modulated carrier. This has been found to be[18]

$$W_1(R,\dot{R},\theta,\dot{\theta};t)=(R^2/4\Omega^2 b_0 b_2)\times\exp[-(R^2+A_0^2-2A_0R\cos(\theta-\chi))/2b_0-(\dot{R}^2+R^2\dot{\theta}^2+A_0^2\dot{\phi}^2+2A_0\dot{\phi}[R\dot{\theta}\sin(\theta-\nu)-\dot{R}\cos(\theta-\nu)])/2b_2], \quad \text{(II.7)}$$

on the assumption of a symmetrical spectral distribution of the noise. Here the angles ϕ and ψ are determined from

$$\cos\chi=\alpha/(\alpha^2+\beta^2)^{\frac{1}{2}};\quad \sin\chi=\beta/(\alpha^2+\beta^2)^{\frac{1}{2}};$$
$$\cos\nu=\dot{\alpha}/(\dot{\alpha}^2+\dot{\beta}^2)^{\frac{1}{2}};\quad \sin\nu=\dot{\beta}/(\dot{\alpha}^2+\dot{\beta}^2)^{\frac{1}{2}}, \quad \text{(II.8)}$$

and the quantities b_n $(n=0,\ 1,\ 2\cdots)$ and Φ are specified by

$$b_n=\int_0^\infty(\omega-\omega_0)^n w(f)df,$$

and

$$\Phi=\omega_d t+\Psi;\quad \dot{\phi}=\omega_d+\dot{\psi}=\omega_d+D_0(t). \quad \text{(II.9)}$$

The expression $w(f)$ is the mean power spectrum of the noise leaving the i-f, so that b_0 is the corresponding mean noise power; the definitions of R, θ, α, and β are given in Section 3.

Integration of (II.7) over $\dot{R}$ and θ gives us at once

$$W_1(R,\dot{\theta};t)=\frac{R^2}{b_0(2\pi b_2)^{\frac{1}{2}}}\exp(-(R^2+A_0^2)/2b_0-(\dot{\theta}^2R^2+A_0^2\dot{\phi}^2)/2b_2)\int_0^{2\pi}(d\theta/2\pi)\times\exp\{A_0^2\dot{\phi}^2\cos^2(\theta-\Psi)/2b_2+A_0R\cos(\theta-\chi)/b_0-A_0\dot{\phi}R\dot{\theta}\sin(\theta-\nu)/b_2\}. \quad \text{(II.10)}$$

Equation (II.10) may then be used to obtain $\langle B_1(R)\dot{\theta}$, according to Eq. (4.1). Simple results follow provided we integrate over θ last. With the aid of the relations

$$\int_{-\infty}^{\infty}\dot{\theta}\exp(-a\dot{\theta}^2\pm b\dot{\theta})d\dot{\theta}=\pm(b(\pi)^{\frac{1}{2}}/2a(a)^{\frac{1}{2}})\exp(b^2/4a),$$

and

$$\int_{-\infty}^{\infty}\dot{\theta}^2\exp(-a\dot{\theta}^2\pm b\dot{\theta})d\dot{\theta}=\frac{1}{2a}\left(\frac{\pi}{a}\right)^{\frac{1}{2}}(1+(b^2/2a))\exp(b^2/4a),\quad a>0 \quad \text{(II.11)}$$

we find

$$\langle B_1\dot{\theta}\rangle=\frac{A_0\dot{\phi}}{b_0}\exp(-A_0^2/2b_0)\int_0^\infty B_1(R)\times\exp(-R^2/2b_0)dR\int_0^{2\pi}\frac{d\theta}{2\pi}\sin(\nu-\theta)\times\exp\left(-\frac{RA_0}{b_0}\cos(\theta-\chi)\right). \quad \text{(II.12)}$$

[30] B. van der Pol, "The fundamental principles of frequency modulation," J. Inst. Elec. Eng. **93**, Part III, 153 (1946).

The integration over θ now offers no difficulties if we use the expansion for the exponential in a series of Bessel functions, *viz.*

$$\exp(A_0R\cos(\theta-\phi)/b_0)$$

$$=\sum_{k=0}^{\infty}\epsilon_k I_k(RA_0/b_0)\cos k(\theta-\chi). \quad \text{(II.13)}$$

The result is

$$I_1(RA_0/b_0)\sin(\phi-\psi),$$

and from (3.4a) $\sin(\phi-\psi)$ is readily shown to be -1. At this point it is convenient to introduce the following normalization:

$$r\equiv R/(2b_0)^{\frac{1}{2}};\quad r_0\equiv R_0/(2b_0)^{\frac{1}{2}};$$
$$a_0\equiv A_0/(2b_0)^{\frac{1}{2}}=(p)^{\frac{1}{2}}\quad(\geq 0). \quad \text{(II.14)}$$

Equation (II.12) may then be written finally

$$\langle E_0(t)\rangle=2\kappa\dot{\phi}a_0\exp(-a_0^2)$$

$$\times\int_0^{\infty}B_1(r)I_1(2a_0r)\exp(-r^2)dr, \quad \text{(II.15)}$$

in which $B_1(r)$ is given by (II.6) on replacing R by $r(2b_0)^{\frac{1}{2}}$, etc., to conform to the normalization (II.14). A similar procedure for $\langle E_0^2\rangle$ gives us

$$\langle E_0(t)^2\rangle=\kappa^2\exp(-A_0^2/2b_0)$$

$$\times\int_0^{\infty}(B_1(R)^2/R)\exp(-R_0^2/2b_0)dR$$

$$\times\int_0^{2\pi}\frac{d\theta}{2\pi}\{\rho^2+A_0^2\dot{\phi}^2\sin^2(\theta-\nu)/b_0\}$$

$$\times\exp(RA_0\cos(\theta-\chi)/b_0),\quad \rho^2\equiv b_2/b_0 \quad \text{(II.16)}$$

corresponding to (II.12). Again integration over θ with the help of (II.13) yields

$$\frac{A_0^2\dot{\phi}^2}{b_0}\{I_0(RA_0/b_0)+I_2(RA_0/b_0)\}+\rho^2I_0(RA_0/b_0),$$

so that we find at last that

$$\langle E_0(t)^2\rangle=\kappa^2\exp(-a_0^2)\Big\{a_0^2\dot{\phi}^2\int_0^{\infty}(B_1(r)^2/r)$$

$$\times\exp(-r^2)[I_0(2a_0r)+I_2(2a_0r)]dr$$

$$+\rho^2\int_0^{\infty}(B_1(r)^2/r)\exp(-r^2)I_0(2a_0r)dr\Big\}. \quad \text{(II.17)}$$

Following substitution of the appropriate expressions for the envelope B_1 obtained from (3.6) or (3.8) we may proceed to determine the mean and mean-square values of the low frequency output of the discriminator as indicated in (II.15) and (II.17). In computation it is more convenient to use the closed form (cf. (3.6)) of the hypergeometric function ${}_2F_1$. Numerical data for $\langle E_0\rangle$ and $\langle E_0^2\rangle$ are illustrated in Figs. 1–3; these data were obtained by numerical integration with the help of Simpson's rule:

$$T=\frac{h}{3}[y_a+y_b+4(y_1+y_3+\cdots)+2(y_2+y_4+\cdots)],$$

where T represents the approximate value of the integral and the y's are the values of the integrand in the interval (a, b), including the end points; h is the (constant) spacing between abscissae. The degree of accuracy of the method is indicated by the difference between the true value T' and its approximation T, which is

$$T'-T\leq Mh^4(b-a)/180,\quad (b>a);$$
$$M=\text{max. value of } y^{(iv)} \text{ in } (a, b).$$

The two cases for which it is possible to effect the integration and so obtain explicit results for $\langle E_0\rangle$ and $\langle E_0^2\rangle$ are (*a*) no limiting ($r_0\to\infty$), and (*b*) *extreme limiting* ($r_0^2\ll 1$, $r_0^2\ll a_0^2$). We examine (*a*) first:

(*a*) Here, as shown in Section 3, the expressions E_0 for Blachman's or the more exact theory, indicated by subscripts B and M respectively, are identical, so that we obtain from Eq. (3.6) ($r_0\to\infty$), substituted into (II.15),

$$\langle E_0(t)\rangle_{M,B}=\eta\kappa\dot{\phi}\left(\frac{\pi b_0}{2}\right)^{\frac{1}{2}}a_0^2\exp(-a_0^2/2)$$

$$\times[I_0(a_0^2/2)+I_1(a_0^2/2)]$$

$$=\eta\kappa\dot{\phi}\left(\frac{\pi b_0}{2}\right)^{\frac{1}{2}}a_0^2\,{}_1F_1(1/2;2;-a_0^2),$$

$$(r_0\to\infty), \quad \text{(II.18)}$$

which is the same as our earlier result, Eq. (2.36) of reference 4 and Blachman's expressions[3] (3.2) and (3.5). When the carrier is weak with respect to the noise, (II.18) reduces to

$$\langle E_0(t)\rangle_{M,B}=\left(\frac{\pi b_0}{2}\right)^{\frac{1}{2}}\eta\kappa\dot{\phi}a_0^2(1-a_0^2/4+a_0^4/16\cdots),$$

$$a_0^2\ll 1,\quad (r_0\to\infty), \quad \text{(II.19)}$$

while for strong carriers we find directly from the asymptotic form of the confluent hypergeometric function ${}_1F_1$

$$\langle E_0(t)\rangle_{M,B}\simeq\eta\kappa\dot{\phi}(2b_0)^{\frac{1}{2}}a_0(1-1/4a_0^2$$

$$-3/32a_0^4-\cdots),\quad a_0^2\gg 1,\quad (r_0\to\infty). \quad \text{(II.20)}$$

Equations (II.18)–(II.20) apply also for finite limiting ($r_0 < \infty$), provided the input carrier-to-noise ratio a_0 is much less than r_0, i.e., $a_0^2 \ll r_0^2$, since then there is essentially no clipping of the carrier or noise because the limiting threshold is rarely exceeded. Corresponding remarks apply for $\langle E_0^2 \rangle$ (cf. (II.21)–(II.23) below). We observe that for weak signals the mean output is proportional to the *square* of the input ratio a_0. This is an example of modulation suppression, in which the relatively strong noise overwhelms the signal, not linearly in a_0, but proportional to a_0^2. Strong carriers, on the other hand, suppress the noise, as we see from (II.20), i.e., $\langle E_0 \rangle \simeq \eta\kappa\phi A_0$, $a_0^2 \gg 1$. A similar dependence of $\langle E_0 \rangle$ on a_0 has been found in a-m reception.[7]

In the same way we obtain from Eq. (3.6) in (II.17)

$$\langle E_0(t)_{M,B}{}^2 \rangle = 2b_0\kappa^2\eta^2 \times \left\{ a_0^2\phi^2(1-(1-\exp(-a_0^2)/2a_0^2)) + \frac{\rho^2}{2} \right\}, \quad (r_0 \to \infty) \quad \text{(II.21)}$$

and this result is precisely Eq. (2.34) of reference 4 (if we note that there $-\phi_2(0) = \rho^2 b_0$ here), or Blachman's expressions (3.7) and (3.8). Again, for weak carriers we find that

$$\langle E_0(t)_{M,B}{}^2 \rangle \doteq b_0\kappa^2\eta^2 \times \{\phi^2 a_0^2(1 + a_0^2/2! - a_0^4/3! + \cdots) + \rho^2\}, \quad a_0^2 \ll 1, \quad (r_0 \to \infty) \quad \text{(II.22)}$$

and for strong ones

$$\langle E_0(t)_{M,B}{}^2 \rangle \simeq 2b_0\kappa^2\eta^2 a_0^2 \left\{ \phi^2\left(1 - \frac{1}{2a_0^2}\right) + \frac{\rho^2}{2a_0^2} \right\}, \quad a_0^2 \gg 1, \quad (r_0 \to \infty). \quad \text{(II.23)}$$

The suppression effects mentioned above for $\langle E_0 \rangle$ are also seen to be in operation here.

(*b*) For extreme limiting we retain only the leading term (in r_0 or r_0^2) and neglect higher powers of r_0. Expansion of the integrands of (II.15) and (II.17), on using Eq. (3.7), allows us to write finally after termwise integration

$$\langle E_0(t) \rangle_M \doteq \frac{4}{\pi}\eta\kappa\phi r_0(2b_0)^{\frac{1}{2}}(1-\exp(-a_0^2)), \quad (r_0^2 \ll 1), \quad \text{(II.24)}$$

and $\langle E_0(t) \rangle_B$ is similar, provided we replace $4/\pi$ above by unity, according to Eq. (3.8). Equation (II.24) is the result (2.37) of reference 4. Weak carriers yield

$$\langle E_0(t) \rangle_M \doteq \frac{4\eta}{\pi}\kappa\phi r_0(2b_0)^{\frac{1}{2}}a_0^2(1 - a_0^2/2! + a_0^4/3! + \cdots), \quad (a_0^2 \ll 1, r_0^2 \ll 1), \quad \text{(II.25)}$$

and for strong carriers we find

$$E_0(t)_M \simeq 4\kappa\eta\phi r_0(2b_0)^{\frac{1}{2}}/\pi = 4\eta\kappa\phi R_0/\pi, \quad (a_0^2 \gg 1, r_0^2 \ll 1). \quad \text{(II.26)}$$

The approximate values (Blachman's theory) for $\langle E_0 \rangle$ are obtained at once from (II.24)–(II.26) on replacing $4/\pi$ by 1. We remark that (II.24), etc., applies equally well for arbitrary limiting, provided the carrier is sufficiently great compared with the clipping level R_0, i.e., $a_0^2 \gg r_0^2$, as then the input wave is limited effectively most of the time, and extreme limiting is still the condition of operation. Suppression of the signal, or of the noise, is again observed, the small-signal behavior being the same with respect to $a_0 (r_0 \ll 1)$ as in the unlimited case. However, very strong carriers now yield an output completely independent of carrier strength and proportional to the limiting level R_0, as we would expect, cf. (II.26).

Extreme values of $\langle E_0^2 \rangle$ may be derived in a similar fashion. We find at length that

$$\langle E_0(t)_M{}^2 \rangle \doteq \frac{16}{\pi^2}\kappa^2\eta^2 r_0^2 b_0 \exp(-a_0^2) \times \{(a_0^2\phi^2 + \rho^2)(\gamma_M - Ei(-r_0^2)) + \sum_{m=0}^{\infty} [\phi^2 a_0^{2m+4}(2m+3) + \rho^2 a_0^{2m+2}(m+2)]/(m+1)(m+2)!\}, \quad r_0^2 \ll 1, \quad \text{(II.27)}$$

where

$$\gamma_M = \pi^2/16 - \sigma_1 = 0.2731$$

and

$$\sigma_1 = 1/3 + 1/90 + 1/5040 - 1681 \cdot 10^{-2}/59584 + \cdots = 0.3444.$$

The quantity σ_1 arises in (II.27) from the terms involving ${}_2F_1{}^2$ (Eqs. (3.6) and (II.17)), if we consider *all* the terms containing r_0^2. The exponential integral $Ei(-r_0^2)$ is defined by

$$Ei(-x) \equiv -\int_x^{\infty} e^{-y}dy/y \doteq (1 - x/2)\log x, \quad x \ll 1.$$

Tables of $Ei(-x)$ are available in Jahnke and Emde. The approximate form of $\langle E_0^2 \rangle$ is

$$\langle E_0(t)_B{}^2\rangle \doteq \kappa^2\eta^2 r_0{}^2 b_0 \exp(-a_0{}^2)$$
$$\times\{(a_0{}^2\phi^2+\rho^2)(1-Ei(-r_0{}^2))$$
$$+\sum_{m=0}^{\infty}[\phi^2 a_0{}^{2m+4}(2m+3)$$
$$+\rho^2(m+2)a_0{}^{2m+2}]/(m+1)(m+2)!\},$$
$$r_0{}^2 \ll 1. \quad \text{(II.28)}$$

The asymptotic values of the series in (II.27) and (II.28) may be found if we express the series in terms of confluent hypergeometric functions ${}_1F_1$ and apply Eq. (A3.3) of reference 6. The results are

$$\sum_{m=0}^{\infty}\frac{a_0{}^{2m+4}(2m+3)}{(m+1)(m+2)!}$$
$$\simeq \exp(a_0{}^2)\{2+1!/a_0{}^2+2!/a_0{}^4+\cdots\},$$

and

$$\sum_{m=0}^{\infty}\frac{a_0{}^{2m+2}}{(m+1)(m+1)!}$$
$$\simeq \frac{\exp(a_0{}^2)}{a_0{}^2}\left(1+\frac{1!}{a_0{}^2}+\frac{2!}{a_0{}^4}+\cdots\right). \quad \text{(II.29)}$$

For weak signals $(\langle E_0{}^2\rangle)_M$ is easily shown from (II.27) to be

$$\langle E_0(t)_M{}^2\rangle \doteq \frac{16}{\pi^2}\kappa^2\eta^2 b_0 r_0{}^2\{\lambda\rho^2+a_0{}^2[\rho^2(1-\lambda)+\lambda\phi^2]$$
$$+a_0{}^4[\rho^2(\lambda/2-3/4)+\phi^2(3/2-\lambda)]+\cdots\},$$
$$a_0{}^2 \ll 1, \quad \text{(II.30)}$$

where $\lambda \equiv (\gamma_M - Ei(-r_0{}^2))$. For $\langle E_0(t)_B{}^2\rangle$ we must multiply (II.30) by $\pi^2/16$ and replace λ by $1-Ei(-r_0{}^2)$. The limiting expressions (II.29) are useful when $a_0{}^2 \gg 1$; in this case we find

$$\langle E_0(t)_M{}^2\rangle \simeq \frac{16}{\pi^2}\eta^2\kappa^2 r_0{}^2 b_0\{\phi^2(2+1!/a_0{}^2+2!/a_0{}^4+\cdots)$$
$$+\rho^2(1+1!/a_0{}^2+2!/a_0{}^4+\cdots)/a_0{}^2\}, \quad \text{(II.31)}$$

and $\langle E_0(t)_B{}^2\rangle$ is identical with (II.31) if we replace $16/\pi^2$ by unity. Observe the suppression of noise in the above, i.e., $2\phi^2$ *vs.* $\rho^2/a_0{}^1$, $a_0{}^2 \to \infty$. Our result (II.31) agrees with Blachman's earlier and less general expression (3.13), reference 3, when $\phi = 0$.

ACKNOWLEDGMENTS

We wish to express our sincere thanks to Miss Marilyn Lang and to Mrs. Roger Stokey, who performed the calculations for this paper.

Effect of Surface Roughness on Eddy Current Losses at Microwave Frequencies

SAMUEL P. MORGAN, JR.
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received October 4, 1948)

A theoretical investigation has been made of the power dissipation by eddy currents in a metallic surface at microwave frequencies in the presence of regular parallel grooves or scratches whose dimensions are comparable to the eddy current skin depth. The eddy current equation has been integrated numerically for grooves of various shapes and sizes transverse to the direction of induced current flow, and the corresponding losses are calculated and plotted. The power dissipation is increased by about 60 percent over its value for a smooth surface when the root-mean-square deviation of the grooved surface from an average plane is equal to the skin depth; the exact shape of the grooves is not critical. The increase in eddy current losses caused by grooves parallel to the current is shown in a particular case to be only about one-third as great as the increase caused by transverse grooves of similar size. The effect on losses of an isolated narrow crack or fissure transverse to the current is briefly discussed.

I. INTRODUCTION

IT is well known that electromagnetic oscillations in a wave guide or cavity resonator are always accompanied by heat losses caused by the flow of induced currents in the metal walls of the structure. The usual formulas for these losses are derived on the assumption of perfectly smooth metal surfaces of conductivity equal to that of the bulk metal. In experimental studies at the Bell Telephone Laboratories, however, it is being found that for wave-lengths in the neighborhood of 3 cm the losses are materially higher than calculated, by amounts ranging roughly from 10 percent to 60 percent. Other work[1] at 1.25 cm has revealed losses similarly in excess of the theoretical values, and the discrepancies tend to become worse at shorter wave-lengths. These discrepancies cannot be explained as attributable to the deviations from Ohm's law which occur[2,3] when the eddy current skin depth becomes less than the mean free path of electrons in the metal, or when the period of the electromagnetic oscillations becomes comparable with the mean time between collisions of an electron with the crystalline lattice, since it now appears that the latter effects will be encountered in good conductors at room temperature only if the wave-length is shorter than a few tenths of a millimeter. We have therefore to seek a classical explanation of the observed increase in losses at centimeter wave-lengths, and the question arises what part the roughness of the metal surface may play.

It has been definitely established[4] that the effective resistance of copper wires at 3300 Mc depends markedly upon the surface treatment to which they are subjected, but the published experimental results which deal directly with the effect of surface roughness on losses are still fragmentary and incomplete. No experimental data are yet available which show quantitatively the dependence of eddy current losses on the degree of roughness of the conducting surface. In the present paper we attempt to give a theoretical discussion of this problem.

At microwave frequencies the induced currents flowing in a conductor are confined to an exceedingly thin surface layer or skin, whose thickness is measured essentially by the characteristic length

$$\delta=(2/\omega\mu g)^{\frac{1}{2}}, \tag{1}$$

where μ and g are the permeability and the conductivity of the metal in m.k.s. units and ω is the angular frequency. For copper a convenient numerical relation is

$$\delta=6.6\nu^{-\frac{1}{2}}\text{ cm},$$

where ν is the frequency in cycles per second. Thus at 10 kc, $\delta=0.066$ cm; at 1 Mc, 0.0066 cm; and at 10^4 Mc, 6.6×10^{-5} cm or 0.66 micron. Otherwise expressed, for 3-cm waves the skin thickness in copper is only 26 microinches, which is of the same order as the roughness of natural finishes such as machined or drawn surfaces. In order to determine to what extent eddy current losses may be increased by surface irregularities whose dimensions are comparable, on the average, to the skin depth, we shall calculate the theoretical magnitude of the effect in a few idealized cases.

The cases treated below are all two-dimensional; i.e., the surface roughness is assumed to consist of infinitely long parallel grooves or scratches either normal to or parallel to the direction of induced current flow. The ratio of power dissipated in a grooved surface to that dissipated in a plane surface under the same external field can be computed by straightforward if laborious numerical methods for grooves of any shape (at least for grooves transverse to the current). The relative power dissipation may be plotted against some standard quantity

[1] E. Maxwell, J. App. Phys. **18**, 629 (1947).
[2] A. B. Pippard, Proc. Roy. Soc. **A191**, 385 (1947).
[3] G. E. H. Reuter and E. H. Sondheimer, Nature **161**, 394 (1948).
[4] C. J. Milner and R. B. Clayton, J. Inst. Elec. Eng. **93**, IIIa, 1409 (1946).

representing the surface roughness,[5,6] for example, the root-mean-square deviation from the mean surface, measured in units of the skin thickness. Since the curves of loss *vs.* roughness are quite similar for grooves of a few simple shapes (e.g., square, rectangular, and equilateral triangular), one may be reasonably confident that the results do not depend critically on the exact shape of the surface. This conclusion naturally breaks down for surfaces which have profiles *widely* different from those studied, or which exhibit flaws such as deep cracks or fissures. In Section IV of this paper a brief approximate treatment is given of the increase in eddy current dissipation resulting from an isolated narrow crack transverse to the current flow.

II. GROOVES TRANSVERSE TO CURRENT FLOW

Consider a semi-infinite conducting solid, as in Fig. 1a or 1b, extending from $y=f(x)$ to $y=-\infty$, where $f(x)$ is a periodic function of period d. This conductor may be part of the surface of a wave guide, coaxial pair, cavity resonator, or microwave antenna; the over-all curvature of the metal surface, if any, will be assumed negligible compared to the reciprocal of the groove width d. Let an alternating magnetic field $H_0e^{i\omega t}$ be impressed in the z direction at the surface $y=f(x)$ so that eddy currents are induced to flow parallel to the xy plane. If the conductor represents the metal surface of a wave guide, cavity resonator, or antenna, H_0 may be computed for the given mode of oscillation on the assumption that the conductivity is infinite, the alteration of the field outside the metal caused by finite conductivity being negligible to first approximation. Since the grooves are very shallow compared to the dimensions of the apparatus, the impressed field does not vary appreciably with x from point to point on the surface. Furthermore the free-space wavelength is so much greater than the depth of penetration of the induced eddy currents that we are justified in neglecting the variation of H_0 with z. Therefore, we shall treat the problem as two-dimensional and H_0 as constant.

In a material of conductivity g, dielectric constant ϵ, and permeability μ, the harmonically varying magnetic field satisfies the eddy current equation

$$\nabla^2\mathbf{H}=i\omega\mu g\mathbf{H}, \tag{2}$$

provided that the ratio $\omega\epsilon/g$ of displacement current to conduction current is negligible compared to unity. This approximation is certainly valid for copper at microwave frequencies.

If we introduce the characteristic skin thickness δ defined by (1) and recall that $\mathbf{H}$ is independent of z and has only a z component, (2) becomes

$$\partial^2H_z/\partial x^2+\partial^2H_z/\partial y^2=(2i/\delta^2)H_z, \tag{3}$$

subject to the boundary conditions $H_z=H_0$ on the surface $y=f(x)$ and $H_z\to 0$ as $|y|\to\infty$.

In the special case where $f(x)\equiv 0$, the surface of the conductor is plane; H_z is independent of x and is simply given by

$$H_z=H_0\exp[-(1+i)|y|/\delta], \tag{4}$$

this being the well-known law of penetration of all harmonically varying field quantities beneath the plane surface of a conductor. If the surface is not plane, we may still expect H_z to vanish exponentially at great depths, but in general it will depend upon both x and y.

Once we have the solution of (3), subject to the assigned boundary conditions, the power dissipated by eddy currents in a given volume V of the conducting material is evidently

$$P=(1/2g)\int_V\mathbf{J}\cdot\mathbf{J}^*dv=(1/2g)\int_V(\nabla\times\mathbf{H})\cdot(\nabla\times\mathbf{H}^*)dv, \tag{5}$$

where $\mathbf{J}$ represents the eddy current density. If we consider a cell of width d equal to the periodic distance in the x direction, unit length in the z direction, and infinite depth, the integral on the right side of (5) may be transformed, as in Appendix I, into

$$P=-(1/g\delta^2)\mathrm{Im}H_0^*\int_0^d\int_{-\infty}^{f(x)}H_z\,dy\,dx, \tag{6}$$

where the integration is extended over the shaded region of Fig. 1a or 1b.

If the surface were plane, from (4) the power dissipated in a strip of unit length and of width d would be just

$$P_0=-\frac{|H_0|^2}{g\delta^2}\mathrm{Im}\int_0^\infty\int_0^d e^{-(1+i)y/\delta}dx\,dy=\frac{|H_0|^2d}{2g\delta}.$$

The ratio of power dissipated per unit length and

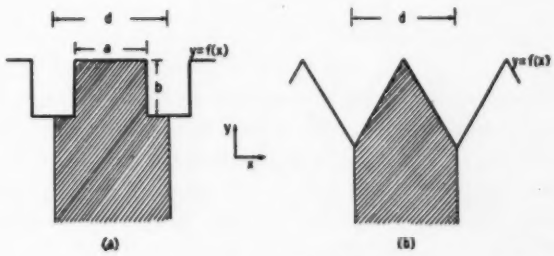


FIG. 1. Cross sections of semi-infinite conductors with rectangular and triangular surface grooves.

[5] *American Standard for Surface Roughness, Waviness, and Lay* (B46.1 American Standards Association, New York, New York, 1947).

[6] R. F. Gagg, E. R. Boynton, and James W. Owens, Industrial Standardization 19, 6 (1948).

width of grooved and smooth surfaces, respectively, is therefore

$$\frac{P}{P_0}=-\frac{2}{|H_0|^2 d\delta}\mathrm{Im}H_0^*\int_0^d\int_{-\infty}^{f(x)}H_z dydx. \quad (7)$$

In order to calculate the relative power dissipation numerically from (7), we have first to solve the partial differential equation (3) with its attendant boundary conditions. An approximate solution giving the values of H_z over a network of points covering the desired region may be obtained with sufficient accuracy for our purpose by the so-called "relaxation method,"[7] details of which are briefly described in Appendix II. Here we shall present merely the results.

Before carrying out numerical analysis we must specify the size and shape of the surface grooves, i.e., the form of the function $f(x)$. The simplest geometrical choices are:

(1) Rectangular:

$$f_1(x)=\begin{cases}0 \text{ for } |x|<\frac{1}{2}a,\\ -b \text{ for } \frac{1}{2}a<|x|<\frac{1}{2}d.\end{cases}$$

(2) Equilateral triangular:

$$f_2(x)=\begin{cases}-\sqrt{3}x \text{ for } 0\leq x\leq\frac{1}{2}d,\\ +\sqrt{3}x \text{ for } 0\geq x\geq-\frac{1}{2}d.\end{cases}$$

Both $f_1(x)$ and $f_2(x)$ are supposed to be periodic in x with the period d.

For the rectangular grooves the mean value of $y_1=f_1(x)$ and the r.m.s. deviation of y_1 from the mean are, respectively,

$$\bar{y}_1=-(d-a)b/d,$$
$$\Delta_1=(b/d)[a(d-a)]^{\frac{1}{2}}.$$

Similarly, for the triangular grooves

$$\bar{y}_2=-\tfrac{1}{4}\sqrt{3}d,$$
$$\Delta_2=\tfrac{1}{4}d.$$

For the actual numerical integration H_0 was assigned the arbitrary value 200 (this value obviously cancels out of Eq. (7) for the power ratio), and the values of H_z were determined to the nearest integer at the lattice points of a regular network covering the strip $0\leq x\leq\frac{1}{2}d$, account being taken of the fact that both types of grooves are symmetric with respect to the planes $x=0$ and $x=\frac{1}{2}d$.

The final values of H_z are recorded on Fig. 2 for a square groove with $a=b=2\delta$, $d=4\delta$, $\Delta_1=\delta$, on a square mesh of interval $\frac{1}{2}\delta$. Having this solution, the double integral in (7) may be evaluated by repeated application of quadrature formulas[8] analogous to Simpson's rule. For the square groove of Fig. 2 we obtain

$$\mathrm{Im}\int_0^d\int_{-\infty}^{f(x)}H_z dydx=-2514\times\tfrac{1}{4}\delta^2.$$

Substitution into (7) gives, since $H_0=200$ and $d=4\delta$,

$$P/P_0=2514/1600=1.57.$$

It may be noted that the solution of this problem on a coarser net with mesh interval δ gave $P/P_0=1.54$, which would probably incline us to trust the above value to within one percent. This is certainly adequate, in view of the fact that we have been treating a highly idealized problem with no immediate prospects for an exact comparison of the theoretical results with experiment.

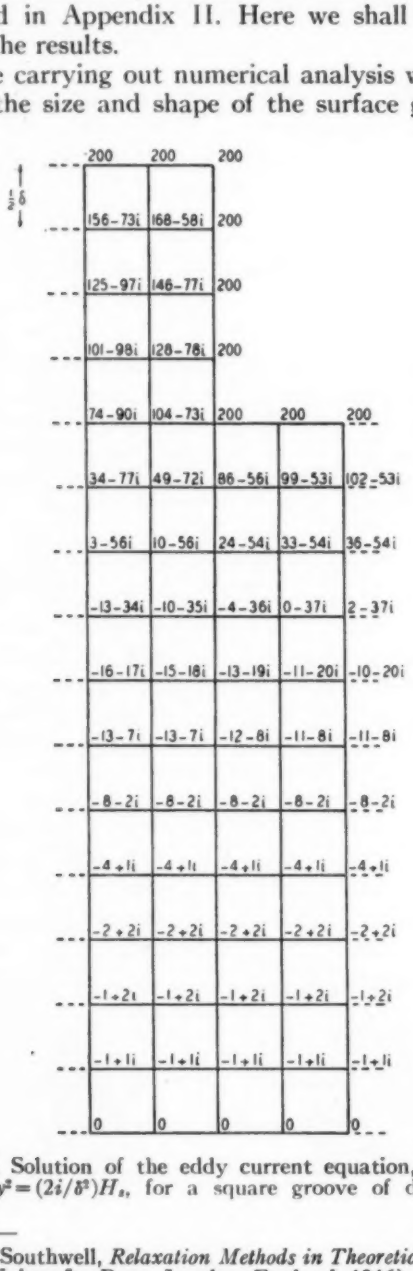



FIG. 2. Solution of the eddy current equation, $\partial^2 H_z/\partial x^2+\partial^2 H_z/\partial y^2=(2i/\delta^2)H_z$, for a square groove of dimensions $2\delta\times 2\delta$.

[7] R. V. Southwell, *Relaxation Methods in Theoretical Physics* (Oxford University Press, London, England, 1946).

[8] *Tables of Lagrangian Interpolation Coefficients* (Columbia University Press, New York, New York, 1944), pp. xxxii–xxxiii.

Calculations exactly similar to those just outlined were carried through for three other sizes of square grooves, for which the ratio $a:b:d=1:1:2$, as well as for three sizes of rectangular grooves for which $a:b:d=3:2:4$. Finally the solution of the eddy current equation (3) was obtained for three sizes of equilateral triangular grooves, the latter calculations being carried out on a triangular-mesh net[9] with a mesh interval equal to $2\delta/3$. All of our numerical results for grooves transverse to current flow are summarized in Table I, which gives the relative power dissipation P/P_0 for different values of the ratio Δ/δ of r.m.s. roughness to skin thickness. These results are plotted in Figs. 3 and 4, where the surface profiles are also drawn to such a scale that the value of Δ is the same for each.

Qualitatively the curves of Figs. 3 and 4 require little interpretation.* It is interesting to consider a few actual numerical magnitudes for 3-cm waves in a copper pipe, where $\delta=26$ microinches. For square grooves with a *peak-to-trough* distance of 20 microinches we have $P/P_0=1.09$; for 30 microinches, 1.23; for 45 microinches, 1.49; and for 90 microinches, 1.74. It was to be anticipated that the relative power dissipation would approach the asymptotic value 2.00 for all three profiles for large values of the ratio Δ/δ, since, when the skin depth is small compared with the dimensions of the grooves, the induced currents follow the surface closely and in each case the total path length through which they flow is doubled over its value for a plane surface. We may expect in general for shallow transverse scratches, whose depth is small compared to the free-space wave-length, that in the limit when the skin depth is small compared to the average dimensions of the scratches the whole surface will be energized uniformly and the losses will merely be proportional to the total exposed area. By what factor the area of an actual metal surface exceeds the area of an ideal smooth surface of the same length and breadth is a matter for experimental rather than theoretical investigation.

III. GROOVES PARALLEL TO CURRENT FLOW

If the impressed magnetic field is parallel to the xy plane of Fig. 1, so that the induced currents all flow in the z direction parallel to the grooves, we have, in general, a much more difficult boundary value problem than that which occurred in Section II. In order to determine the electromagnetic field or the current density within the conductor, we need to know the tangential component H_t of the external magnetic field at the surface. For example, in terms of the current density J_z it is easy to show that the continuity requirement on H_t at the surface leads (at least if the trace of the surface in the xy plane consists of straight-line segments) to the boundary condition

$$\partial J_z/\partial n=(2i/\delta^2)H_t,$$

where $\mathbf{n}$, $\mathbf{t}$, $\mathbf{k}$ are a right-handed system of unit vectors specifying, respectively, the outward normal to the conductor, the tangent to the conductor in the xy plane, and the direction of the z axis.

Unfortunately, we do not generally know the tangential component of the surface field, and we cannot even make the convenient assumption which was possible in Section II, namely, that H_t is approximately independent of frequency. Clearly, when the frequency is zero or the conductivity is zero, the lines of H run straight across the grooves and ridges in Fig. 1, completely unaffected by the presence of the metal. Of course, no eddy currents are induced by a rigorously static field (or in a perfect dielectric). When the frequency is infinite or the conductivity is infinite the field does not penetrate at all into the metal. The boundary is a line of H, and the surface field varies markedly from point to point, being strong near sharp corners and weak at the bottoms of grooves. In the general case of finite conductivity and finite frequency, the lines of H are gradually forced out of the conductor as the frequency rises, the extent to which this occurs depending on the ratio of the dimensions of the grooves to the skin depth δ.

An exact solution of the problem of eddy current losses in grooves parallel to the current flow would require the simultaneous determination of the electromagnetic fields both outside and inside the conductor. While such an exact solution has not been carried out, it is possible to compute the external field on the assumption that the metal surface is a perfect conductor, and from this to obtain a good

TABLE I. Relative power dissipation *vs.* root-mean-square roughness for surface grooves transverse to current flow.

Δ/δ	P/P_0 Square grooves	Rectangular grooves	Triangular grooves
0	1.00	1.00	1.00
0.25	1.04	—	—
0.43	—	1.25	—
0.50	1.17	—	1.24
0.87	—	1.60	—
1.00	1.57	—	1.61
1.52	—	1.72	—
1.67	—	—	1.80
1.75	1.75	—	—

[9] Reference 7, pp. 11, 20–24, *et passim*. The use of the greater mesh interval is probably justified by the fact, here pointed out, that an inherently more accurate finite difference approximation can be employed with a triangular mesh than with a square mesh.

* It is probably unnecessary to remark that slight apparent implausibilities in the relative shapes of the plotted curves, such as the more pronounced elbow near $\Delta/\delta=1.00$ of the lower curve in Fig. 4, are doubtless insignificant; they could easily be due to the uncertainty in the third figure of our values for P/P_0.

approximation to the eddy current losses when the conductivity is finite but the transverse dimensions of the grooves are large compared to the characteristic skin depth. These losses may be compared with the losses caused by grooves of similar size and shape transverse to the current, for which the power dissipation, by the remarks at the end of Section II, is merely proportional to the total surface area. In this way we may get some indication of the relative effect of the two kinds of grooves on surface losses.

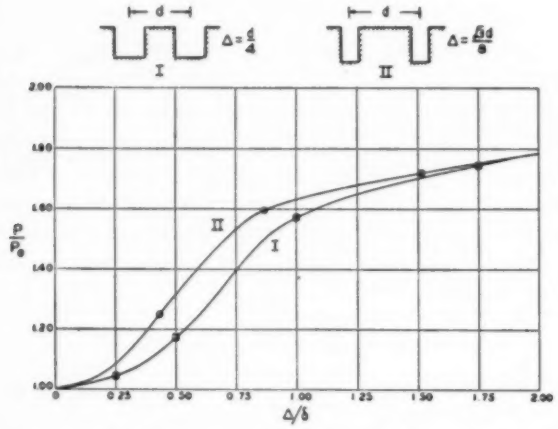


FIG. 3. Relative power dissipation P/P_0 *vs.* root-mean-square roughness Δ/δ for square and rectangular grooves transverse to current flow.

We wish, accordingly, to calculate the field distribution when an originally uniform alternating magnetic field is bounded on one side by a perfectly conducting surface having regular parallel grooves or ridges transverse to the direction of the field, the dimensions of the grooves being much smaller than the free-space wave-length corresponding to the given frequency. Advantage may be taken of the fact that, when the wave-length is sufficiently long compared with the dimensions of the region under observation, the instantaneous field configurations are indistinguishable from those given by the static solution of Laplace's equation which satisfies the same boundary conditions. To see this, we observe that the free-space propagation equation may be written in the form

$$\nabla^2\mathbf{H}=(2\pi/\lambda)^2\mathbf{H}, \tag{8}$$

where λ is the free-space wave-length. Suppose that $\mathbf{H}$ varies an appreciable fraction of its total value within a distance d; more precisely, assume that some of the second derivatives on the left side of (8) are comparable in magnitude with H/d^2. Now from (8),

$$d^2\nabla^2\mathbf{H}=4\pi^2(d/\lambda)^2\mathbf{H},$$

so that, if $d/\lambda\ll 1$, we commit a negligible error by replacing (8) with

$$\nabla^2\mathbf{H}=0. \tag{9}$$

A harmonically varying field which approximately satisfies (9) will be called "quasi-static."

A two-dimensional quasi-static magnetic field $\mathbf{H}e^{i\omega t}$ may be derived from a stream function[10] $\Psi(x, y)e^{i\omega t}$ by

$$H_x=\partial\Psi/\partial y, \quad H_y=-\partial\Psi/\partial x, \quad H_z=0, \tag{10}$$

where Ψ satisfies the two-dimensional Laplace equation

$$\partial^2\Psi/\partial x^2+\partial^2\Psi/\partial y^2=0. \tag{11}$$

On a perfectly conducting boundary the normal component of $\mathbf{H}$ vanishes and the stream function Ψ is constant. Since (11) is satisfied by either the real or the imaginary part of any analytic function W of the complex variable $\zeta=x+iy$, our problem reduces to finding a suitable complex stream function,

$$W=\Phi+i\Psi, \tag{12}$$

whose imaginary part is constant on conducting boundaries. For a grooved or ridged boundary whose trace on the ζ-plane consists entirely of straight-line segments, W may be found by a Schwarz transformation.[10] The transformation giving the rectangular grooves of Fig. 1a is derived in terms of elliptic functions in Appendix III.

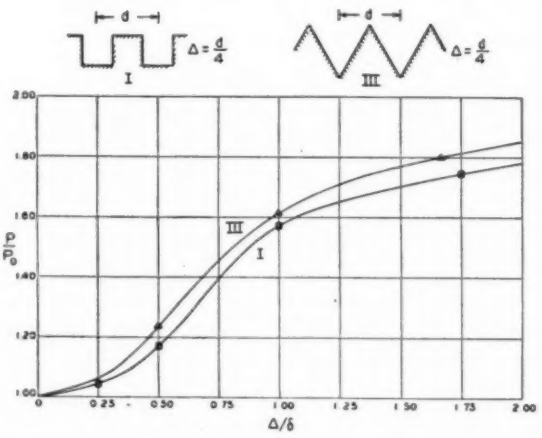


FIG. 4. Relative power dissipation P/P_0 *vs.* root-mean-square roughness Δ/δ for square and equilateral triangular grooves transverse to current flow.

The power dissipated by eddy currents in a surface S of conductivity g, whose radius of curvature is large compared to the skin depth δ, is given approximately by[11]

$$P=(1/2g\delta)\int_S |H_t|^2 ds, \tag{13}$$

[10] S. A. Schelkunoff, *Electromagnetic Waves* (D. van Nostrand Company, Inc., New York, New York, 1943), pp. 174, 179–180, 184 ff.

[11] Reference 10, p. 320.

where H_t is the amplitude of the tangential magnetic field impressed at the conducting surface. For the grooved surface under consideration, if the skin depth is very small compared to the transverse dimensions of a typical groove, we may obtain a good approximation to the dissipated power by using in (13) the value of H_t which has been calculated on the assumption of a perfectly conducting surface. Of course (13) does not apply at a corner where the radius of curvature is zero; but if the size of the grooves is large compared to the skin depth only a small fraction of the total power will be dissipated in the immediate vicinity of sharp corners.

To the approximation considered, the integral in (13) is evaluated for the case of rectangular grooves in Appendix III. The ratio of eddy current power spent in a groove of width d to the power spent in a plane strip of width d and equal length is found to be

$$\frac{P}{P_0}=\frac{2K'}{\pi}\left[\frac{k'^2\mathrm{sn}(h,k')\mathrm{cn}(h,k')}{\mathrm{dn}(h,k')}\right]+1-\frac{2b}{d}, \quad (14)$$

where the quantities h and k' are related to the dimensions a, b, d of Fig. 1a by the following pair of equations:

$$h=(a/d)K'+(2b/d)K, \quad (15)$$

$$2Z(h,k')-\frac{2k'^2\mathrm{sn}(h,k')\mathrm{cn}(h,k')}{\mathrm{dn}(h,k')}+\frac{\pi(2b/d)}{K'}=0. \quad (16)$$

In these equations K and K' are the complete elliptic integrals of the first kind to the complementary moduli k and k', respectively, and $Z(h,k')$ is the zeta-function of Jacobi.[12] Thus we see that, whereas the limiting value of P/P_0 for grooves transverse to the current and large compared to the skin depth is just equal to the relative increase in total surface area, the limiting value for rectangular grooves parallel to the current is a function of the root of a pair of complicated transcendental equations involving the dimensions of the grooves.

Numerical results have been obtained only for grooves of square cross section, i.e., grooves with

$$a=b=\tfrac{1}{2}d.$$

In this case Eqs. (15) and (16) become

$$h=K+K'/2,$$

$$Z(h,k')-\frac{k'^2\mathrm{sn}(h,k')\mathrm{cn}(h,k')}{\mathrm{dn}(h,k')}+\frac{\pi}{2K'}=0.$$

[12] E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge University Press, London, England, 1940), fourth edition, p. 518.

These equations were solved for k by a trial and error process using a table of elliptic functions;[13] the quantities of interest were found to be

$$k=\sin 3°16'=0.05698,\quad k'=\sin 86°44'=0.99838,$$
$$K=1.5721,\quad K'=4.2545,\quad h=3.6993.$$

Substitution of these numerical values into (14) gives for the limiting ratio of eddy current losses resulting from square grooves parallel to the current:

$$(P/P_0)_{\parallel}=1.360. \quad (17)$$

Since the corresponding ratio for square grooves transverse to the current is

$$(P/P_0)_{\perp}=2.000, \quad (18)$$

the increase in power dissipation due to large "parallel" grooves is only about one-third as much as the increase due to large "perpendicular" grooves in this case. Whether the curves representing $(P/P_0)_{\parallel}$ *vs.* Δ/δ would be completely similar in shape to the curves of Figs. 3 and 4 for $(P/P_0)_{\perp}$ of course we cannot be sure, but from the comparative asymptotic values (17) and (18) it seems reasonable to believe that, insofar as anomalous eddy current losses may be attributable to surface grooves and scratches, transverse grooves have a considerably greater adverse effect than grooves parallel to the current.

IV. ISOLATED DEEP CRACK TRANSVERSE TO CURRENT

In addition to shallow scratches it is quite likely that a metal surface which has been strained in the process of working will exhibit occasional narrow cracks and fissures whose depth may be several times the skin thickness, though still small compared with the free-space wave-length corresponding to the given frequency. An approximation to the eddy current power dissipated in such a crack transverse to the current may be obtained by regarding it as a section of parallel-plane transmission line.

Consider an infinitely long rectangular groove of width b and length l (Fig. 5) in a semi-infinite conducting medium of electrical constants ϵ, μ, g. Neglecting edge effects near the opening, unit length of this groove is equivalent to a transmission line[14] for transverse magnetic waves (H perpendicular to plane of paper) whose distributed series impedance and shunt admittance per unit distance

[13] E. P. Adams and R. L. Hippisley, *Smithsonian Mathematical Formulae and Tables of Elliptic Functions* (Smithsonian Institution, Washington, D. C., 1922), pp. 260–309. Since we have used linear interpolation uncritically with these tables, the numerical values in this section are probably given to one more figure than is significant.

[14] Reference 10, pp. 484 ff.

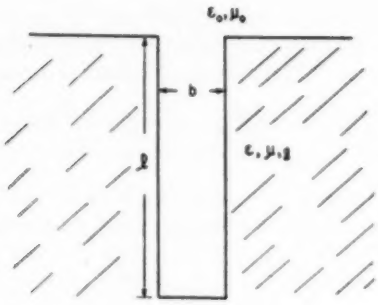


FIG. 5. Cross section of isolated rectangular groove in a conducting surface.

are, respectively,

$$Z=2\eta+i\omega L=2(1+i)/g\delta+i\omega\mu_0 b, \tag{19}$$

$$Y=i\omega C=i\omega\epsilon_0/b. \tag{20}$$

Here ϵ_0, μ_0 are the dielectric constant and permeability of free space, δ is the skin thickness defined by (1), and

$$\eta=(i\omega\mu/g)^{\frac{1}{2}}=(1+i)/g\delta \tag{21}$$

is the intrinsic impedance of the conducting material at the given angular frequency ω.

From familiar transmission line theory[15] we know that if the line is terminated in an impedance

$$Z_l=\eta b=b(1+i)/g\delta,$$

the input impedance will be

$$Z_i=K\frac{Z_l\,\mathrm{ch}\Gamma l+K\,\mathrm{sh}\Gamma l}{K\,\mathrm{ch}\Gamma l+Z_l\,\mathrm{sh}\Gamma l}, \tag{22}$$

where

$$\Gamma=(YZ)^{\frac{1}{2}},\quad K=(Z/Y)^{\frac{1}{2}}. \tag{23}$$

Thus the ratio of the power dissipated by a current I flowing across the groove to the power dissipated in the absence of the groove by the same current flowing across a flat strip of width b is

$$\frac{P}{P_0}=\frac{\mathrm{Re}\tfrac{1}{2}Z_i|I|^2}{\mathrm{Re}\tfrac{1}{2}\eta b|I|^2}=\frac{\mathrm{Re}Z_i}{(b/g\delta)}. \tag{24}$$

In the special case where the width of the groove is comparable to or greater than the skin thickness and the depth is much less than the free-space wave-length, i.e., $b/\delta\approx 1$ and $l/\lambda\ll 1$, we have from (19), (20), and (23), on writing $\lambda=(2\pi/\omega)(\mu_0\epsilon_0)^{-\frac{1}{2}}$, the relation

$$|\Gamma l|=(2\pi l/\lambda)|(1+\delta/b)-i\delta/b|^{\frac{1}{2}}\approx(2\pi l/\lambda)\ll 1.$$

Hence $\mathrm{sh}\Gamma l\approx\Gamma l$ and $\mathrm{ch}\Gamma l\approx 1$, and substituting these approximations into (22) we get, after some manipulation,

$$Z_i\approx\frac{(1/g)[(b+2l)/\delta+i(b/\delta+2l/\delta+2lb/\delta^2)]}{1+\frac{1}{2}(i-1)(2\pi/\lambda)^2 l\delta}.$$

The second term in the denominator is negligible compared with unity under the given assumptions. Hence, from (24) the ratio of power dissipated in the groove to power dissipated on the plane surface is simply

$$P/P_0=(b+2l)/b. \tag{25}$$

Thus the increase in eddy current loss is proportional to the additional surface area introduced by the walls of the groove, provided $b\approx\delta$.

For a very narrow crack it is to be expected that the shunt capacitance between the walls will become the dominant effect and will ultimately short-circuit the crack. To investigate this we assume now that $b/\delta\ll 1$ and obtain the following approximations:

$$Z_l=(b/\delta)(1+i)/g,$$
$$\Gamma l\approx(2\pi l/\lambda)(i-1)^{\frac{1}{2}}(b/\delta)^{-\frac{1}{2}},$$
$$K\approx -i(\lambda/\pi g\delta)(i-1)^{\frac{1}{2}}(b/\delta)^{\frac{1}{2}},$$
$$\mathrm{ch}\Gamma l\approx\tfrac{1}{2}\exp[(2\pi l/\lambda)(i-1)^{\frac{1}{2}}(b/\delta)^{-\frac{1}{2}}]\approx\mathrm{sh}\Gamma l,$$
$$Z_i\approx K\approx(1.099-0.455i)(\lambda/\pi g\delta)(b/\delta)^{\frac{1}{2}}.$$

The input impedance of a groove of fixed depth thus vanishes like $(b/\delta)^{\frac{1}{2}}$ when the ratio $b/\delta\to 0$, but not so fast as the impedance of a plane strip of width b, which vanishes directly as b/δ. If b/δ is so small that the last approximation for Z_i holds, the ratio of powers is given by (24) to be

$$P/P_0=1.099(\lambda/\pi\delta)(b/\delta)^{-\frac{1}{2}}. \tag{26}$$

V. CONCLUSIONS

The effect of shallow parallel scratches on the surface of a conductor is to increase eddy current losses by an appreciable amount if the dimensions of the scratches are comparable on the average to the skin depth. If the scratches are transverse to the direction of induced current flow the increase in loss may be anywhere between zero and approximately 100 percent, depending on the ratio of r.m.s. surface roughness to skin depth. This increase does not depend critically on the exact shape of the surface profile, so long as the scratches are of approximately equal width and depth.

The adverse effect on eddy current losses of scratches parallel to the current flow is considerably less than the adverse effect caused by transverse scratches of similar size. Our analysis suggests that the increase in relative power dissipation resulting from "parallel" grooves or scratches may be only slightly more than one-third as great as the increase in the transverse case.

The eddy current dissipation in an isolated narrow crack transverse to the current is approxi-

[15] Reference 10, Chapter VII.

mately proportional to the total surface area of the walls, provided that the depth of the crack is much less than the free-space wave-length corresponding to the given frequency and the width is comparable to or greater than the skin thickness. For an extremely narrow crack the shunt capacitance between the walls becomes important, and the losses are ultimately proportional to the square root of the ratio of width to skin thickness.

ACKNOWLEDGMENTS

The author wishes to thank Dr. S. A. Schelkunoff for numerous helpful suggestions. He has also enjoyed stimulating conversations at various stages of the work with Dr. W. A. Tyrrell, who originally suggested the investigation.

The numerical calculations of Section II were carried out by Miss J. G. Asbury.

APPENDIX I

Transformation of the integral

$$\int_V (\nabla\times\mathbf{H})\cdot(\nabla\times\mathbf{H}^*)dv.$$

In the integral

$$2gP=\int_V (\nabla\times\mathbf{H})\cdot(\nabla\times\mathbf{H}^*)dv, \qquad \text{(A1)}$$

write

$$\nabla\times\mathbf{H}=\nabla\times H_z\mathbf{k}=\nabla H_z\times\mathbf{k},$$

where $\mathbf{k}$ is the unit vector in the z direction. Then,

$$2gP=\int_V (\nabla H_z)\cdot(\nabla H_z^*)dv,$$

since ∇H_z is normal to $\mathbf{k}$. On applying Green's theorem,

$$\int_V \nabla U\cdot\nabla W dv=\int_S W(\partial U/\partial n)ds-\int_V W\nabla^2 U dv,$$

first with $U=H_z$ and $W=H_z^*$, then with $U=H_z^*$ and $W=H_z$, we obtain

$$2gP=\tfrac{1}{2}\Big[\int_S (H_z^*(\partial H_z/\partial n)+H_z(\partial H_z^*/\partial n))ds$$

$$-\int_V (H_z^*\nabla^2 H_z+H_z\nabla^2 H_z^*)dv\Big], \qquad \text{(A2)}$$

where n is the outward normal to the surface S bounding the volume V. In virtue of (3) and its complex conjugate equation,

$$\nabla^2 H_z^*=-(2i/\delta^2)H_z^*,$$

the volume integral in (A2) vanishes. If the integration is extended over a cell of unit length and of width d, the surface integrals over the ends $z=$ constant vanish because $\partial H_z/\partial z=0$, and the contributions of the side surfaces $x=x_0$ and $x=x_0+d$ cancel each other on account of periodicity. The fields vanish exponentially as $|y|\to\infty$, so we are left with only the integral over the external surface S_1 where $H_z=H_0=$ constant. Hence

$$2gP=\mathrm{Re}\int_{S_1} H_z^*(\partial H_z/\partial n)ds=\mathrm{Re}H_0^*\int_{S_1}(\partial H_z/\partial n)ds.$$

The last integral may be re-extended over the entire surface S of the volume under consideration since, as before, the net contribution of the sides and the surface at infinity will be zero. Once more applying Green's theorem, this time with $U=H_z$ and $W=1$, and then Eq. (3), we finally get

$$2gP=\mathrm{Re}H_0^*\int_S(\partial H_z/\partial n)ds=\mathrm{Re}H_0^*\int_V\nabla^2 H_z dv$$

$$=\mathrm{Re}H_0^*\int_V(2i/\delta^2)H_z dv$$

$$=-(2/\delta^2)\mathrm{Im}H_0^*\int_0^d\int_{-\infty}^{f(x)} H_z dy dx,$$

which is Eq. (6) of Section II.

APPENDIX II

Relaxation method applied to the eddy current equation. We seek a numerical solution of the equation

$$\partial^2 w/\partial x^2+\partial^2 w/\partial y^2=(2i/\delta^2)w, \qquad \text{(A3)}$$

which takes the constant value $w=W_0$ on the boundary $y=f(x)$, is symmetric with respect to the lines $x=0$ and $x=\frac{1}{2}d$, and vanishes as $y\to-\infty$.

The first step is to replace the partial derivatives in (A3) by finite differences. Let w_0 be the value of w at an interior point of a square-mesh net such as that of Fig. 2, and let w_1, w_2, w_3, w_4 be the surrounding values, as in Fig. 6. Then it may be shown[16] that the finite-difference approximation to (A3) is

$$\tfrac{1}{4}\sum_{n=1}^{4} w_n-w_0[1+(ih^2/2\delta^2)]=0, \qquad \text{(A4)}$$

up to terms of order h^4, where h is the mesh interval.

The relaxation method of solving (A4) consists in the following procedure: Having assumed over the given net a solution, however crude, which

[16] Reference 7, p. 20.

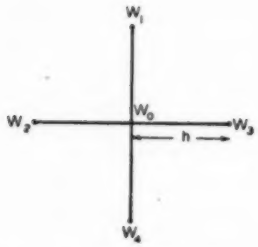


FIG. 6. Typical point of a "relaxation net."

satisfies the boundary conditions, we define at each interior point a *residual* R proportional to the left side of (A4). By altering the value of w_0 at a given point we can reduce the value of the residual at that point, at the expense of simultaneous changes in the residuals at the immediately adjacent points; successive corrections at points where the residuals are largest gradually bring us closer to the desired solution. Theoretically there exists a set of w's for which every residual is zero; in practice, however, we adopt some limit of tolerance and regard the solution as complete when all residuals are reduced below this value.

As an example, to obtain the solution of Fig. 2 we take a mesh interval $h=\frac{1}{2}\delta$ in (A4). Multiplying the left side of (A4) by 8 to clear of fractions and writing $w=u+iv$, we define the residual to be

$$R=(2\sum_{n=1}^{4}u_n-8u_0+v_0)+i(2\sum_{n=1}^{4}v_n-u_0-8v_0). \quad \text{(A5)}$$

We then assume, using as an aid to guessing any physical insight or *a priori* knowledge of the nature of the solution that we may possess, a set of values of w and compute the initial residuals by (A5), noting that if w_0 lies on one of the lateral boundaries of symmetry, then $w_2=w_3$. It is convenient to decide in advance the number of decimals which will be wanted in the solution and then to multiply the boundary value W_0 by such a factor that the actual calculations may be done in integers. Thereafter the values of w and the residuals are altered according to the following scheme, derived from (A5):

Δw_0	ΔR_0	$\Delta R_{1,2,3,4}$
$+1$	$-8-i$	$+2$
$+1$	$+1-8i$	$+2i$

If w_2 (or w_3) lies on a boundary of symmetry, then the corresponding ΔR is $+4$ or $+4i$, respectively, because the change in w_0 is reflected on the other side of the boundary.

The degree to which liquidation of residuals can ultimately be carried depends on the minimum correction we are willing to make in w; for example, if we round off w to the nearest integer we should accept residuals whose real and imaginary parts separately do not exceed 4 units, if they are defined by (A5). This limit of tolerance was imposed in obtaining the solution of Fig. 2. When a solution is deemed acceptable, one checks the work by recomputing all residuals directly from the final values of w.

An empirical check on the accuracy of our results for the square groove of dimensions $2\delta\times2\delta$ was provided by carrying out the complete solution on two different nets. The mesh intervals employed were $h=\delta$ and $h=\frac{1}{2}\delta$; the corresponding values of the power ratio P/P_0 were 1.54 for the coarse net and 1.57 for the fine net. A mesh interval of $\frac{1}{2}\delta$ was then used for all the rectangular grooves except the square one of dimensions $\frac{1}{2}\delta\times\frac{1}{2}\delta$, in which case the interval was $\frac{1}{4}\delta$.

For the finite-difference approximation analogous to (A4) which may be used with a triangular-mesh net, reference should be had to a text[9] on relaxation methods.

APPENDIX III

Conformal representation of a boundary with rectangular grooves. Starting with the complex stream function

$$W=H_0\zeta, \quad \text{(A6)}$$

which represents a uniform field H_0 parallel to the x axis, we subject the ζ-plane to the following conformal transformations (Fig. 7):

$\zeta_1=\zeta_1(\zeta)$ takes a semi-infinite vertical strip of width $2a$ in the ζ-plane into the entire upper half of the ζ_1-plane.

$\zeta_2=\zeta_2(\zeta_1)$ takes the upper half of the ζ_1-plane into a semi-infinite vertical strip of width $2a_2$ in the ζ_2-plane, with a symmetrical rectangular notch of height b_2 and width $2c_2$ in the lower boundary.

Let it be assumed that the following correspondences exist between pairs of points in Fig. 7, where for reasons of symmetry the notation differs somewhat from that employed in Fig. 1a and in the main body of this paper:

$$0\rightleftarrows0_1\rightleftarrows0_2, \quad \text{(A7)}$$

$$c\rightleftarrows c_1\rightleftarrows c_2, \quad \text{(A8)}$$

$$b\rightleftarrows b_1\rightleftarrows c_2-ib_2, \quad \text{(A9)}$$

$$a\rightleftarrows a_1\rightleftarrows a_2-ib_2. \quad \text{(A10)}$$

Then the required transformations are, in differential form:

$$d\zeta/d\zeta_1=C_1(\zeta_1^2-a_1^2)^{-\frac{1}{2}}, \quad \text{(A11)}$$

$$d\zeta_2/d\zeta_1=C_2(\zeta_1^2-a_1^2)^{-\frac{1}{2}}(\zeta_1^2-b_1^2)^{-\frac{1}{2}}(\zeta_1^2-c_1^2)^{\frac{1}{2}}, \quad \text{(A12)}$$

and hence

$$d\zeta_2/d\zeta=C(\zeta_1^2-b_1^2)^{-\frac{1}{2}}(\zeta_1^2-c_1^2)^{\frac{1}{2}}, \quad \text{(A13)}$$

where C_1, C_2, and $C=C_2/C_1$ are as yet arbitrary

constants. From (A11) and (A13) and the left members of (A7)–(A10) we easily obtain

$$\zeta_1 = a_1 \sin(\pi\zeta/2a)$$

and

$$\zeta_2 = \int_0^{\zeta} \left[\frac{\sin^2(\pi\zeta/2a) - \sin^2(\pi c/2a)}{\sin^2(\pi\zeta/2a) - \sin^2(\pi b/2a)}\right]^{\frac{1}{2}} d\zeta, \quad \text{(A14)}$$

the constant C in (A13) having been put equal to unity so that we shall have $d\zeta_2/d\zeta \to 1$ when $\mathrm{Im}\zeta$ is large and positive. The constants a, b, and c have ultimately to be evaluated in terms of a_2, b_2, and c_2, which specify the dimensions of the groove; evidently $a = a_2$ if $C = 1$.

To simplify writing put

$$\theta = \pi\zeta/2a, \quad \beta = \pi b/2a, \quad \gamma = \pi c/2a. \quad \text{(A15)}$$

The substitutions

$$t = \tan\theta/\tan\gamma, \quad k = \tan\gamma/\tan\beta \quad \text{(A16)}$$

give

$$\zeta_2 = (2a/\pi)\int_0^{\theta} [(\sin^2\theta - \sin^2\gamma)/(\sin^2\theta - \sin^2\beta)]^{\frac{1}{2}} d\theta$$

$$= \frac{2a}{\pi}\frac{\sin^2\gamma}{\sin\beta\cos\gamma}\left[\int_0^t \frac{dt}{[(1-t^2)(1-k^2t^2)]^{\frac{1}{2}}}\right.$$

$$\left. - \sec^2\gamma \int_0^t \frac{t^2 dt}{(1+t^2\tan^2\gamma)[(1-t^2)(1-k^2t^2)]^{\frac{1}{2}}}\right]. \quad \text{(A17)}$$

Let

$$t = \mathrm{sn}(u, k), \quad u = \mathrm{sn}^{-1}(t, k), \quad du = [(1-t^2)(1-k^2t^2)]^{-\frac{1}{2}} dt. \quad \text{(A18)}$$

Then

$$\zeta_2 = \frac{2a}{\pi}\frac{\sin^2\gamma}{\sin\beta\cos\gamma}\left[u - \sec^2\gamma\int_0^u \frac{\mathrm{sn}^2 u\, du}{1+\tan^2\gamma\,\mathrm{sn}^2 u}\right]. \quad \text{(A19)}$$

The last integral is an elliptic integral of the third kind,[17] which will be in the standard form if we write

$$\tan^2\gamma = -k^2\mathrm{sn}^2(ih, k), \quad \text{(A20)}$$

i.e., recalling (A16),

$$\mathrm{sn}(ih, k) = i\tan\beta, \quad \mathrm{cn}(ih, k) = \sec\beta, \quad \mathrm{dn}(ih, k) = \sec\gamma. \quad \text{(A21)}$$

Jacobi's imaginary transformation[18] gives

$$\mathrm{sn}(h, k') = \sin\beta, \quad \mathrm{cn}(h, k') = \cos\beta, \quad \mathrm{dn}(h, k') = \cos\beta\sec\gamma, \quad \text{(A22)}$$

where

$$k^2 + k'^2 = 1.$$

From (A22), since $0 < \beta < \pi/2$, h is real and

$$0 < h = \mathrm{am}^{-1}(\beta, k') < K'.$$

[17] Reference 12, pp. 522–523.
[18] Reference 12, pp. 505–506.

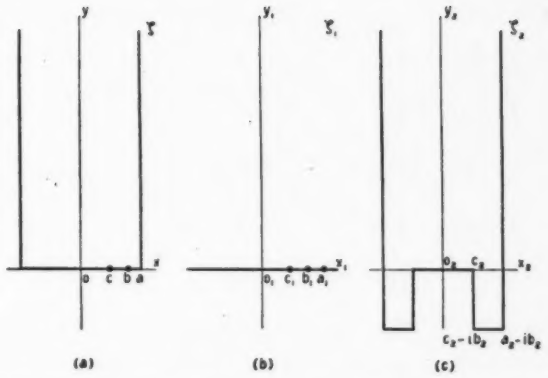


FIG. 7. Conformal transformation of a plane boundary into a boundary with regular rectangular grooves.

Now

$$\int_0^u \frac{\mathrm{sn}^2 u\, du}{1 - k^2\mathrm{sn}^2(ih)\mathrm{sn}^2 u} = \frac{1}{k^2\mathrm{sn}(ih)\mathrm{cn}(ih)\mathrm{dn}(ih)} \times\left[\tfrac{1}{2}\log\frac{\Theta(u-ih)}{\Theta(u-ih)} + uZ(ih)\right], \quad \text{(A23)}$$

where the modulus of the elliptic functions is understood to be k unless otherwise indicated. It will be convenient to use the relation[19]

$$Z(ih, k) = i\left[\frac{\mathrm{dn}(h, k')\mathrm{sn}(h, k')}{\mathrm{cn}(h, k')} - Z(h, k') - \frac{\pi h}{2KK'}\right]. \quad \text{(A24)}$$

Substituting (A23) and (A24) into (A19) and employing (A21) and (A22) to replace the trigonometric functions of β and γ with functions of h and k, we obtain after considerable reduction

$$\zeta_2 = \frac{a}{\pi}\left[Au + i\log\frac{\Theta(u-ih)}{\Theta(u+ih)}\right], \quad \text{(A25)}$$

where

$$A = 2Z(h, k') + \frac{\pi h}{KK'} - \frac{2k'^2\mathrm{sn}(h, k')\mathrm{cn}(h, k')}{\mathrm{dn}(h, k')}. \quad \text{(A26)}$$

As ζ varies from 0 to a, t varies from 0 to ∞, and u from 0 to K to $K+iK'$ to iK'. In tracing the variation of ζ_2 we must be careful to stay on the same branch of the logarithmic function. If we do so, we find the correspondences in the following

[19] Reference 12, p. 519.

table:

ζ	t	u	ζ_2
0	0	0	0
c	1	K	aAK/π
b	$1/k$	$K+iK'$	$aA[K+iK']/\pi-aih/K$
a	∞	iK'	$aAiK'/\pi-aih/K+a$

If we now introduce the constants a_2, b_2, c_2 which specify the dimensions of a typical groove in Fig. 7c, we see from (A8)–(A10) that the following relations must hold:

$$aAK/\pi = c_2, \tag{A27}$$

$$ah/K - aAK'/\pi = b_2, \tag{A28}$$

$$a = a_2. \tag{A29}$$

The scale factors a and a_2 are equal, as previously noted; h and k depend on the ratios c_2/a_2 and b_2/a_2. Elimination of A from (A27) and (A28) gives

$$h=(c_2/a_2)K'+(b_2/a_2)K; \tag{A30}$$

this value of h may be substituted into the expression for A given by (A29). Thus, employing (A26) with (A27),

$$2Z(h,k')-\frac{2k'^2\mathrm{sn}(h,k')\mathrm{cn}(h,k')}{\mathrm{dn}(h,k')}+\frac{\pi(b_2/a_2)}{K'}=0. \tag{A31}$$

Comparing Fig. 7c with Fig. 1a, we see that the quantities a_2, b_2, c_2 correspond to $\frac{1}{2}d$, b, $\frac{1}{2}a$, respectively, so that (A30) and (A31) are equivalent to (15) and (16) of Section III.

If Ψ and W are defined by Eqs. (10) and (12) of Section III, it follows from the Cauchy-Riemann relations[10] that the magnitude of the field represented by W at any point of the ζ_2-plane is just

$$H=|\partial W/\partial\zeta_2|=|H_0|\,|\partial\zeta/\partial\zeta_2|.$$

Employing Eq. (13) to calculate the relative power dissipation in the grooved surface, we have

$$\frac{P}{P_0}=\left[\int_0^{a_2}|H_0|^2|d\zeta/d\zeta_2|^2|d\zeta_2|\right]\Big/\left[\int_0^a|H_0|^2|d\zeta|\right]$$

$$=\frac{1}{a}\int_0^a\left|\frac{\sin^2(\pi\zeta/2a)-\sin^2(\pi b/2a)}{\sin^2(\pi\zeta/2a)-\sin^2(\pi c/2a)}\right|^{\frac{1}{2}}|d\zeta|, \tag{A32}$$

if we make use of (A14). We introduce again the substitutions (A15) and (A16) and obtain

$$\frac{\dot{P}}{P_0}=\frac{2\sin\beta}{\pi\cos\gamma}\int_0^\infty\left|1-\frac{t^2\tan^2\gamma\csc^2\beta}{1+t^2\tan^2\gamma}\right|\times\frac{dt}{|(1-t^2)(1-k^2t^2)|^{\frac{1}{2}}}. \tag{A33}$$

In order to make use of our previous work, we shall find it convenient to have the abbreviations

$$U_{p,q}\equiv\int_p^q\frac{dt}{|(1-t^2)(1-k^2t^2)|^{\frac{1}{2}}},$$

$$V_{p,q}\equiv\int_p^q\frac{t^2dt}{(1+t^2\tan^2\gamma)|(1-t^2)(1-k^2t^2)|^{\frac{1}{2}}}.$$

It follows from (A17) and from the table below (A26) that

$$U_{0,1}=K,\quad V_{0,1}=\left[\cos^2\gamma-\frac{A\sin\beta\cos\gamma}{2\tan^2\gamma}\right]K;$$

$$U_{1,1/k}=K',$$

$$V_{1,1/k}=\left[\cos^2\gamma-\frac{A\sin\beta\cos\gamma}{2\tan^2\gamma}\right]K'+\frac{\pi h\sin\beta\cos\gamma}{2K\tan^2\gamma};$$

$$U_{1/k,\infty}=K,$$

$$V_{1/k,\infty}=\left[\cos^2\gamma-\frac{A\sin\beta\cos\gamma}{2\tan^2\gamma}\right]K+\frac{\pi\sin\beta\cos\gamma}{2\tan^2\gamma}.$$

Hence from (A33)

$$\frac{P}{P_0}=\frac{2}{\pi}\frac{\sin\beta}{\cos\gamma}\left[U_{0,1/k}-\frac{\tan^2\gamma}{\sin^2\beta}V_{0,1/k}-U_{1/k,\infty}+\frac{\tan^2\gamma}{\sin^2\beta}V_{1/k,\infty}\right]$$

$$=\frac{2K'}{\pi}\left[\frac{\sin^2\beta-\sin^2\gamma}{\sin\beta\cos\gamma}\right]+\frac{AK'}{\pi}-\frac{h}{K}+1.$$

If we substitute for A from (A27) and for h from (A30), the expression for P/P_0 simplifies to

$$\frac{P}{P_0}=\frac{2K'}{\pi}\left[\frac{\sin^2\beta-\sin^2\gamma}{\sin\beta\cos\gamma}\right]+1-\frac{b_2}{a_2},$$

which is equivalent, by (A16) and (A22), to Eq. (14) of Section III.

Flash Radiography Applied to Ordnance Problems*

J. C. CLARK**
Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland
(Received September 7, 1948)

Flash radiographic techniques as applied to Army Ordnance problems are discussed. The feasibility of using microsecond x-ray bursts in the study of internal ballistic and explosive phenomena is discussed and illustrated by selected flash radiographs.

I. INTRODUCTION

THE analysis of many phenomena involving rapid motion has often been simplified by the application of flash, spark, and very high speed camera photographic techniques. However, there are many instances when optical methods cannot be applied, either due to the phenomena being shielded by optically opaque material, or accompanied by an intense light flash sufficient to obliterate the effect under study. In such cases it is generally possible to apply the flash radiographic technique. In general, flash radiography implies a burst of x-rays of over-all time duration of the order of a microsecond, and for practical purposes the x-ray burst must be sufficiently intense, and with effective wave-length, to form a satisfactory radiographic image.

Various authors[1–6] have described or reported flash radiographic apparatus involving special types of x-ray tubes and accompanying electrical circuits which produce bursts of x-rays varying in time of duration, intensity, and effective wave-length. A common feature of all flash x-ray apparatus described to date involves the discharge of a high voltage capacitor through the x-ray tube, although the principle of operation of the tube may vary. Of the various authors, only Slack and Ehrke developed a "sealed-off," commercial, high current tube—that is, one which need not be attached to a pumping system or used with a refrigerant during its operation. This cold cathode, field emission type tube is manufactured by Westinghouse Electric and Manufacturing Company and can be operated at potentials as high as 350 kv in air, or higher when oil immersed. The auxiliary equipment which has been developed by Westinghouse[7] for use with this tube consists of a surge generator utilizing the conventional Marx circuit, together with pulsed initiating and timing circuits for controlling the instant the x-ray burst occurs. Because the x-ray tube currents which flow may be 3000 or more amperes, the target is made of tungsten, and, to prevent excessive evaporation, the focal spot is not restricted to the small size most desirable for radiographic purposes. This sacrifice is made in radiographic definition in favor of tube-life, normally 500–1000 flashes.

Since their development, two of the Westinghouse Micronex units have been in use at the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, as well as several other Army and Navy laboratories. It is the purpose of this paper to describe some of the applications of the flash radiographic technique to Ordnance research problems considered at Aberdeen.

II. NATURE OF THE X-RAY BURST

From the blur of the radiographic image produced by a rapidly moving rifle bullet, Slack[3] determined that the duration of the x-ray flash from his equipment was of the order of magnitude of one microsecond. Kingdon and Tanis[1] measured the duration

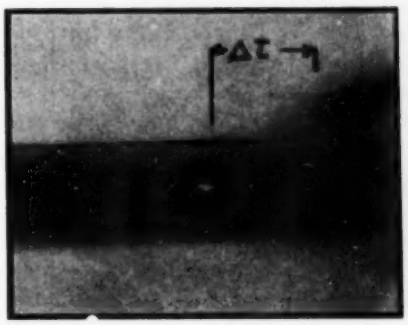


FIG. 1. Flash radiograph of lead indicator strip to show nature of x-ray burst. Detonation of pentolite charge moved from right to left. The x-ray burst occurred during the time indicated by Δt, and the nature of the x-ray burst is shown by the multiple images produced by the indicator strip.

* This work was performed for the Research and Development Division of the U. S. Ordnance Department. The experiments discussed in Section 2, Section 3B(4), and 3B(5) were performed in collaboration with the Eastern Laboratory of the Explosives Department of E. I. du Pont de Nemours and Company, under Contract W-670-ORD-4331. Dr. L. B. Seely (now at Los Alamos Scientific Laboratory) of the Eastern Laboratory actively participated and contributed in a great measure in this research.

** Now at Los Alamos Scientific Laboratory.

[1] K. H. Kingdon and H. E. Tanis, Phys. Rev. **53**, 128 (1938).

[2] M. Steenbech, *Wissenschaftliche Veröffentlichen aus den Siemens Werken* **17**, p. 363 (1938), Chapter 4.

[3] C. M. Slack and L. F. Ehrke, J. App. Phys. **12**, 165 (1941).

[4] W. J. Oesterkampf, Phillips Tech. Rev., 22 (1940).

[5] V. A. Zuckerman, Comptes Rendus (Doklady) de l'Academie des Sciences de l'URSS **XL**, 267 (1943); also Russian J. Tech. Phys. **12**, 185 (1942).

[6] L. E. Simon, *German Research in World War II* (John Wiley and Sons, Inc., New York, 1947), pp. 37, 114, 119.

[7] Trade name of Westinghouse equipment is "Micronex."

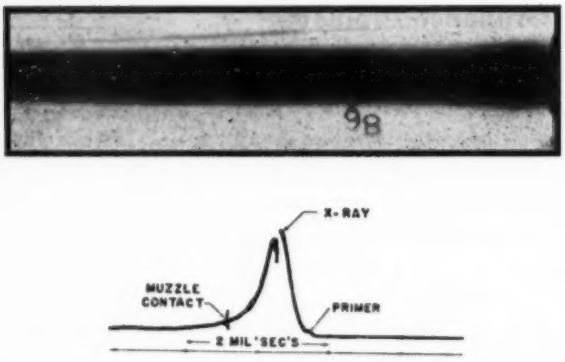


FIG. 2. (a) Flash radiograph of caliber 0.30 projectile moving in rifle barrel. Radiograph taken at instant pressure is maximum. (b). Oscillograph record of pressure *versus* time. The instant the x-ray burst occurred is shown.

of the x-ray burst from the length of the radiographic image of a stationary point source produced on a film fastened to a rapidly rotating disk. It is seen, however, that if a reasonable time resolution of an x-ray burst lasting only one or two microseconds is desired, it becomes necessary to move either the film or the indicator producing an image on the film with a velocity at least 5000 meters per second.

The latter method has been employed by the author. This entails making a flash radiograph of an indicator consisting of a thin strip of lead foil cemented to a small piece of explosive of high detonation velocity, as the detonation proceeds along the piece of explosive. Such a flash radiograph is shown in Fig. 1. In this case a 0.002-in. thick lead foil strip $\frac{1}{8}$-in. wide was cemented along the top of a stick of cast pentolite $\frac{1}{2}$-in. in diameter. The detonation progressed from right to left at the rate of 7.4 mm per microsecond, and the undisturbed lead strip along the charge indicates the section of the explosive which detonated during the x-ray flash. This flash radiograph illustrates, in general, the duration and nature of the x-ray bursts produced. The exposures vary, but for the case illustrated it is seen that the flash consisted of a multiple discharge, producing four x-ray bursts, each variable in intensity and duration and extending over a total time of two microseconds.

The above described technique for analyzing the nature of short duration bursts of x-rays is simple and direct, involving the use of only small explosive charges, and the problem of blast shielding both film and x-ray apparatus is not a difficult one.

III. SOME APPLICATIONS OF FLASH RADIOGRAPHIC TECHNIQUE

A. Internal Ballistics

Provided the intensity and effective wave-length of the microsecond burst of x-rays are sufficient to penetrate the thickness of steel involved, this technique can be used for projectile motion within a gun barrel. Also, since flash photographs of the projectile emerging from the muzzle of the gun are often obliterated by the muzzle flash, the radiographic technique can be applied to muzzle emergence studies.

For internal ballistics studies on guns, piezocrystal pressure gauges are used to obtain an oscillographic record of the instantaneous values of the pressure within the gun barrel established by the burning propellant. It is possible to indicate upon this pressure-time record the instant the x-ray flash occurs, and from the flash radiograph showing the position of the projectile in the gun and the pressure-time curve, travel-time data can be obtained. Figure 2(a) is a flash radiograph of a caliber 0.30 projectile moving in the barrel at the instant the burning propellant has established maximum pressure behind the projectile, as shown from the $P-T$ curve in Fig. 2(b). The projectile in this case has moved approximately its own length from the rest position. By use of two or more of the flash radiographic units, properly timed, successive positions of each projectile during its motion along the barrel can be obtained.

Figure 3 shows a flash radiograph made of a 90-mm projectile emerging from the muzzle of the gun at approximately 2600 ft./sec. The projectile in this case was an experimental model upon which a study was made on methods of attaching the windshield. As shown by the radiograph, the method of attachment of the windshield to the projectile was defective in this particular case.

Figure 4 is a photograph of the arrangement of the flash radiographic equipment about the gun during the field tests. The surge generator and x-ray tube are mounted in the trailer, protected from the muzzle blast by an armor plate shielding, and the x-ray film holder and shielding are opposite the

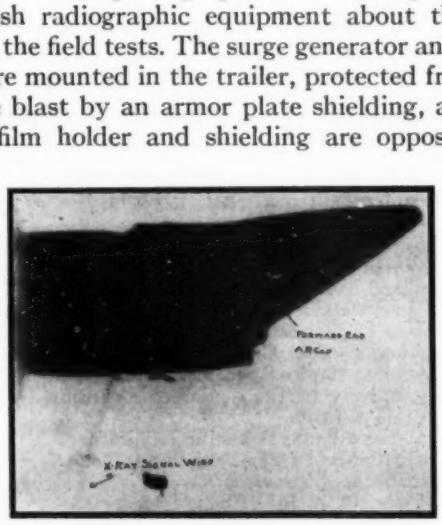


FIG. 3. Flash radiograph made in connection with experiments on methods of attaching windshield to 90-mm projectile. In this case the projectile is emerging from the muzzle brake of the gun at approximately 2600 ft./sec., and the method of attachment of windshield is defective.

x-ray tube. For such field studies a portable darkroom and power source were used to expedite the tests.

Still another application of this technique was the determination of the behavior of rocket propellants at various stages of burning, with and without the rockets in flight. Figure 5 shows a flash radiograph of rocket propellant during an early stage of burning, and with the rocket in flight. The rocket is also spinning (for stabilization) in this case. The propellant consisted of six sticks of 1-in. O.D. approximately 8-in. long with a hole along the axis of the stick for mounting on a wire, squirrel-cage type support. The twisting of the sticks of propellant due to the rotational forces can be seen.

B. Explosive Phenomena

(1) *Constancy of Detonation of Primacord*

Perhaps the simplest high explosive material to use for laboratory studies is Primacord.[8] This consists of approximately a $\frac{1}{8}$-in. diameter core of powdered PETN firmly packed in a cloth casing of approximately $\frac{1}{4}$-in. O.D. Primacord can be initiated with a No. 6 electric blasting cap. The detonation velocity for Primacord is very nearly 6.2 mm per microsecond and quite constant over lengths of several meters. For this reason it can be used to an advantage for controlling and timing explosive events.

The uniformity of detonation of Primacord is illustrated in Fig. 6, which shows a flash radiograph made during the detonation of four pieces of Primacord which were simultaneously initiated by butting the initiating ends against a common Tetryl pellet. Indicator strips, consisting of a 0.002-in. thick lead foil, $\frac{1}{8}$-in. wide, were cemented along the outside of the four pieces of Primacord to more easily determine the progress of the detonation in each piece. The lengths of Primacord from point of initiation to position of detonation head at the instant the radiograph was taken were slightly more than 13 inches. It is seen that a very small variation in detonation velocity occurred in the four pieces.

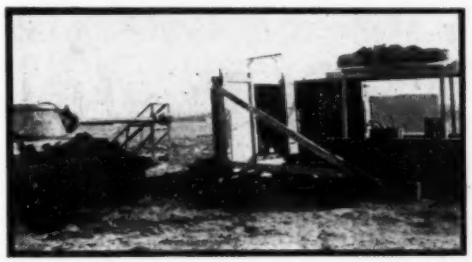


FIG. 4. Flash radiographic equipment, blast shielding, and film holder arranged about 90-mm gun in field emplacement.

[8] Primacord is manufactured by the Ensign Bickford Company, Hartford, Connecticut.

FIG. 5. Flash radiograph of burning rocket propellant in motor of 4.5-in. rocket during flight. This was a spinner rocket, and the twisting of the propellant sticks may be seen.

(2) *Primacord Switch and Radiographic Timing*

The instant the burst of x-rays occurs in the Micronex equipment is controlled by the initiation of the surge generator circuit. The surge generator circuit is initiated by the application of a potential pulse of approximately 15 kilovolts to a pair of spark gaps, critically adjusted. This pulse is produced by discharging a capacitor charged to a few kv through the primary of a steep rise potential transformer, by means of a suitable thyratron. If the FG-27A thyratron is used, it is desirable to pulse the grid with at least 800 volts in order to reduce the variation in conduction time. For many problems involving flash radiographic studies of motion as fast as that associated with explosive phenomena, the following technique was found[9] very suitable for controlling the close tolerance in timing which was required. The technique involves placing two electrodes (household pins), separated about 1 mm, into the PETN core of the Primacord, and connecting these into a circuit as shown in Fig. 7. Normally the PETN is non-conducting, but immediately after its detonation the explosion gases are highly conductive. This arrangement thus serves as a switch to initiate the thyratron pulsing circuit used for controlling the surge generator. In addition, if another piece of Primacord of known length is used to initiate the explosive charge, and both explosive charge and timing Primacord are radiographed by the same x-ray burst, the exact time at which the radiograph is made can be determined.

Use was made of the above Primacord clock technique for timing and controlling the flash radiographs shown in Figs. 10–13.

(3) *Fragmentation Patterns*

(*a*) *Electric blasting caps.*—An interesting example of the fragmentation produced by a cased high explosive charge is that from a commercial

[9] This method of timing flash radiographs was also developed independently during the same period by foreign investigators.

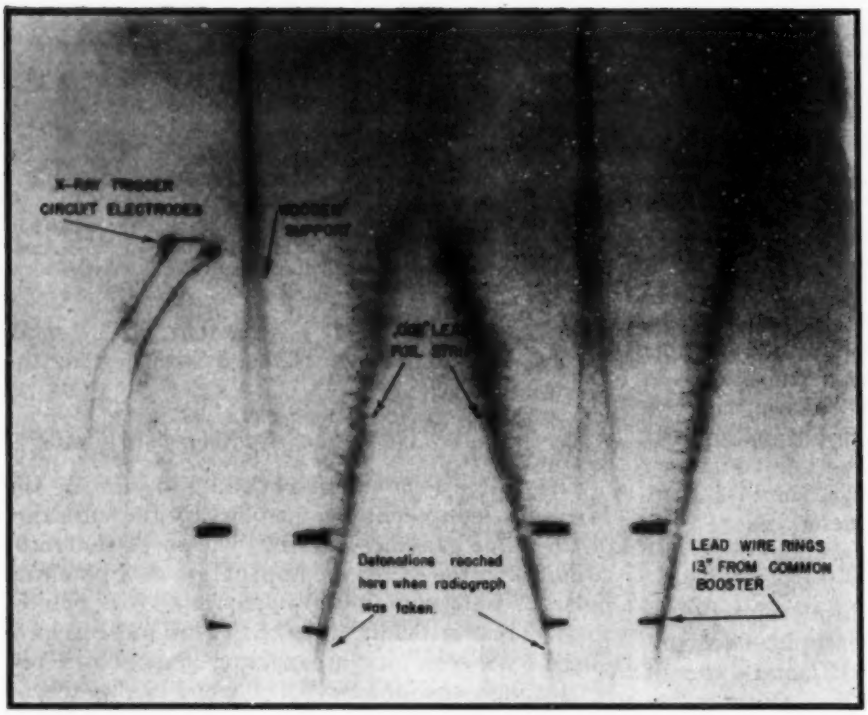

X-RAY TRIGGER CIRCUIT ELECTRODES
Detonations reached here when radiograph was taken.
LEAD WIRE RINGS 12" FROM COMMON BOOSTER

FIG. 6. Flash radiograph of four pieces of Primacord simultaneously initiated, to show constancy of detonation of this material.

electric blasting cap. This type of cap consists of a high resistance bridge wire surrounded by a bead of heat-sensitive explosive, such as lead azide, followed by a main charge of a more stable explosive, and the entire assembly encased in a few mil thick copper shell. The forward end of the shell is often indented, forming a small cavity in the charge. The cap is initiated by sending a current through the bridge wire. Figs. 8(a)–(d) show a series of flash radiographs made of the fragmentation patterns produced by No. 6 electric blasting caps.[10] Only one flash radiograph was made of each cap during its detonation, but the time each radiograph was made was varied to furnish the series shown. A pin, placed in close proximity to the expanding detonator shell, served as an electrical contactor to initiate the x-ray discharge.

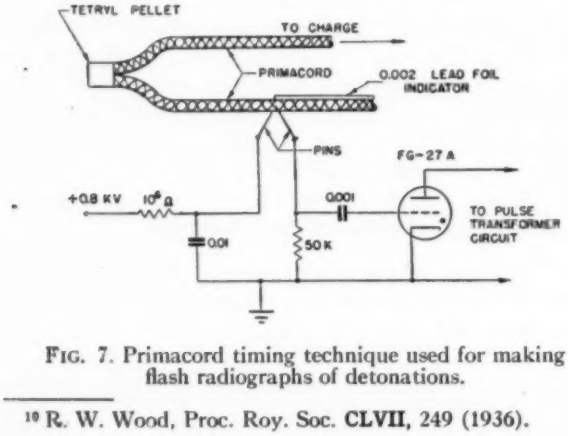


FIG. 7. Primacord timing technique used for making flash radiographs of detonations.

[10] R. W. Wood, Proc. Roy. Soc. **CLVII**, 249 (1936).

The effect of the cavitation at the forward end of the shell of the detonator can be seen in Fig. 8(e). This is a flash radiograph of a No. 8 electric blasting cap. The No. 8 cap is similar to the No. 6 shown above, except that it contains more explosive. In this radiograph a clearly defined metal jet[11] and slug appear. This jet of copper, produced by the collapsing copper shell in the region of the cavity, is observed to consist of a well-defined beam of very fast moving case material. When this jet is directed against an aluminum target, cavitation of the target, characteristic of such jets, results. The effect of depth of penetration in the target as well as hole size depends upon the so-called "stand-off" distance (i.e., cavity-target distance), as shown in the photographs of Fig. 9. The above-described metal jets from the electric blasting caps are exactly the same as those produced in the much larger hollow charges used in the warhead of the now famous bazooka weapon.

(*b*) *20-mm H.E. shells.*—Using the technique for radiographic timing described in Fig. 7, the fragmentation patterns produced by metal-cased high explosive charges during detonation have been studied. Figures 10(a)–(d) show such a series of flash radiographs made of the 20-mm H.E. shell during and immediately after static detonation.[12] By carefully timing these radiographs, using the

[11] C. Munroe, Pop. Sci. Monthly **56**, 444 (1900).

[12] A very interesting study of similar shell upon target impact during flight has been made at Frankford Arsenal by E. R. Thilo, Proc. Nat. Electronics Conference held at Chicago, October 5, 1944.

Primacord clock technique described above, determinations were made of initial fragment velocity for different parts of the fragmentation pattern.

(*c*) *Ring type charges.*—Dispersion studies in the radial velocity of fragments following static detonation of a metal-encased high explosive charge were also made with specially designed charges. One such type of charge assembly, shown in Fig. 11(á), consists of a steel ring 2 mm thick and 4.8 mm long fitted snugly over a 12.7-mm O.D. cast pentolite charge 50 mm long. The detonation was initiated in the charge at the end farthest away from the steel ring so that the detonation would be well established by the time it reached the ring. Flash radiographs of this charge assembly were made with the charge axis parallel and perpendicular to the x-ray film. The results are shown in Figs. 11(b)–(d). Figure 11(b) shows the fragmentation pattern produced by the ring 14.0 microseconds after the detonation passed the ring, and with the axis of the charge perpendicular to the x-ray film. Two distinct rings of fragments are observed. Figure 11(c) shows the fragmentation from the same type of charge, 11.9 microseconds after the detonation passed the ring and with the axis of the charge parallel to the film. It is seen that the ring with the smaller fragments comes from the forward end of the charge and that these fragments have a forward component of velocity. As the detonation strikes the metallic ring a strong induced shock wave is established in the steel ring. Upon reflection of the induced shock from the forward end of the ring, stresses are set up which greatly exceed the elastic properties of the ring [Hopkinson bar effect[13]], and spalling occurs. If a copper ring of the same dimensions is used, results as shown in Fig. 11(d) are obtained. Here two rings of fragments forward and one backward are spalled off the ring. Presumably because of the ductility of copper, fragmentation of the central portion of the ring occurs later than for steel, in fact, late enough for shock reflections from both ends to occur. When thicker steel rings are used spalling occurs from both the forward and rear ends of the ring, probably as a result of fragmentation of the central section of the ring occurring late enough for the reflected shock from the front end of the ring to travel undisturbed to the back end of the ring.

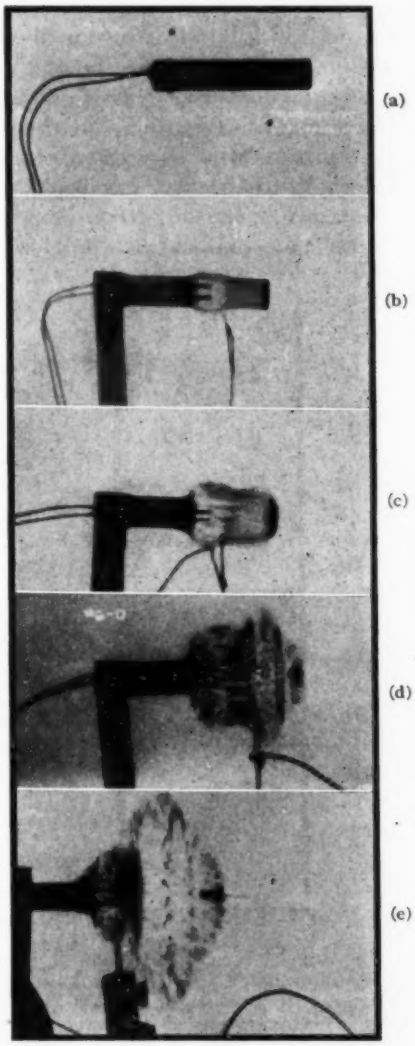


FIG. 8. Flash radiographs of No. 6 and No. 8 electric blasting caps. (a) No. 6 cap before detonation; (b) No. 6 cap during detonation; (c) No. 6 cap, detonation complete. Fragmentation of copper case has occurred; (d) fragmentation pattern produced by No. 6 cap several microseconds after detonation; (e) fragmentation pattern produced by No. 8 cap several microseconds after detonation. The metal jet produced from this cap can be seen moving to the right along the axis of the charge.

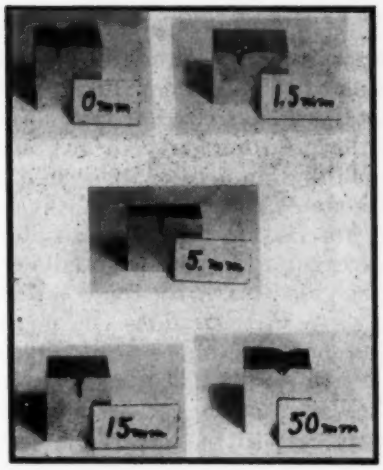
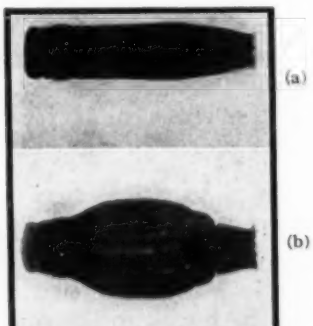


FIG. 9. Photographs of the cavitation produced in aluminum targets by the metal jets from No. 8 electric blasting caps, at the "stand-off" distances indicated.

[13] B. Hopkinson, Phil. Trans. Roy. Soc. London **A213**, 437 (1914).

(4) *Particle Motion in Detonations*

The shock wave in a detonating explosive mixture, maintained by the accompanying rapid chemical reaction, is characterized, in part, by a discontinuity in the pressure, temperature, and density of the detonating material as compared with that in the undetonated state. Following the detonation wave is a rarefaction wave. Flash radiography offers an excellent technique for studying the particle motions associated with the compression and rarefaction waves in detonations. One method which has been employed for solid opaque explosives consists of inserting indicators of 0.0002-in. thick lead foil disks between ½-in. thick disks of explosive. The explosive was cast so that the ends of the disks were very smooth. Several such disks were then stacked together with the lead foil indicators between them, and the entire charge assembly pressed together and securely cemented along the outside. In addition, a 0.002-in. thick lead foil indicator strip was cemented along the top of the assembled charge for the purpose of accurately determining the position of the detonation wave at the instant the radiograph was made. A "zero" radiograph of the charge assembly was made before detonation to serve as a reference picture. A radiograph was next made of the charge during detonation. By comparing the two radiographs the displacement of the thin indicator disks was determined, and from several such pairs of radiographs, for exactly similar charge assemblies, enough data was collected to determine a complete history of the motion of the indicators. The 0.0002-in. thick lead foil indicator is equivalent in mass to a disk of the explosive material only 0.0015-in. thick, and the assumptions are made that the lead foil indicator has no effect upon the detonation and that the motion of the lead disk represents to a good approximation the particle motion of the explosive material in the detonation and rarefaction waves. In Figs. 12(a)–(c) are shown flash radiographs of a high explosive charge as-



(d)

FIG. 10. Flash radiographs of 20-mm H.E. shell (a) At instant detonation of high explosive reached base of shell. Relatively thick shell case has started to expand; (b) approximately 16 microseconds after 10(a); (c) approximately 25 microseconds after 10(a); (d) approximately 54 microseconds after 10(a). The dotted line indicates undetonated shell.

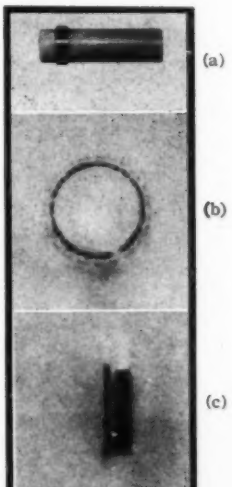


FIG. 11. Ring-type charges. (a) Photograph of 2-mm thick steel ring on pentolite charge ½-in. diameter and 2-in. long. The charge is detonated from the right end; (b) flash radiograph of fragmentation pattern produced by charge assembly shown in 11(a). The charge was mounted with its axis perpendicular to the x-ray film (parallel to the x-ray beam), and the radiograph was made 14 microseconds after detonation passed the ring; (c) flash radiograph of fragmentation pattern produced by charge assembly shown in 11(a). The charge was mounted with its axis parallel to the x-ray film and the radiograph made 11.9 microseconds after detonation passed the ring; (d) flash radiograph of fragmentation pattern produced by 2-mm thick copper ring-type charge, similar to 11(a). The charge axis was parallel to the x-ray film, and the radiograph made 12.5 microseconds after the detonation passed the ring.

sembly as described above, before (Fig. 12(a)) and during (Fig. 12(b)) detonation. The lead strip indicator along the top of the charge shows the detonation wave to have reached a position halfway between the last two disks. By comparing Figs. 12(a) and 12(b), as shown in 12(c), the particle motion can be seen to be forward (compression) near the detonation and later backwards (rarefaction) a few microseconds after the detonation has passed.

(5) *Jets from Explosive Charges with Metal-Lined Cavities*

A cavity or indentation in the otherwise smooth surface of an explosive charge has long been known to influence greatly the blast wave produced in the cavity region by the detonation. Although the effect was described earlier abroad, Charles Munroe[11] appears to be the first investigator in this country to have reported experiments with cavities in high explosive charge assemblies. A lined cavity charge consists of an explosive charge assembly with the cavity surrounded by a liner, generally of some suitable metal. Typical of the cavity charges suitable for producing so-called "Munroe jets" of metal were those used in the warhead of the bazooka weapon. This type of charge carried a conical steel cavity liner.

In order to understand better the phenomena of cavity collapse and jet formation, flash radiographic studies were first made of relatively small size cavity charges with a 45° conical cavity and bearing a corresponding 45° steel liner. The charges were placed as close to the x-ray film as possible, with suitable blast shield both between the charge and film and between the charge and x-ray equipment. Timing of the x-ray flash was accomplished by the Primacord clock technique already described, and only one flash radiograph was made of each charge either during or immediately after detonation. By varying the radiographic timing of several identical charges, a complete history of the jet formation phenomena was obtained. Figure 13(a) shows a radiograph of one of the charges before detonation, and Fig. 13(b) shows the charge at the instant the detonation has just reached the flange at the end of the 45° conical steel liner. The collapsing liner can easily be seen,[14] with motion being directed to the axis of the charge. Also in this radiograph may be seen the first stages of jet formation along the axis of the charge where part of the collapsing liner has reached the axis. Figure 13(c) is a radiograph taken at a later stage of liner collapse, a few microseconds after all of the explosive had detonated. Figure 13(d) is a radiograph made approximately 22 microseconds after all explosive had detonated, and shows the resulting metal jet produced by this type of charge.

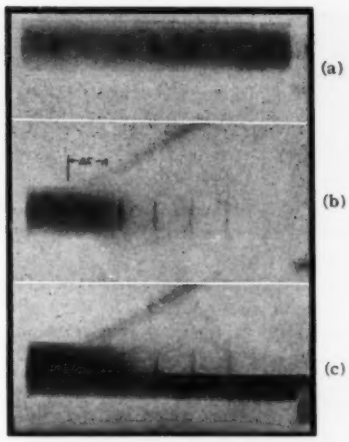


FIG. 12. Particle motion in detonations. (a) Radiograph of charge assembly used for studying particle motion in the region of the detonation. The 0.0002-in. thick lead indicator disks are seen along the axis of the charge, and the 0.002-in. thick lead indicator strip is along the top of the charge; (b) flash radiograph of charge assembly shown in 12(a), during detonation. During the x-ray burst the detonation progressed from right to left over a distance indicated by Δt; (c) radiographs 12(a) and 12(b) compared. The lower half of the picture is 12(a), the upper half 12(b), and the displacement of the 0.0002-in. thick lead disks indicates the displacement of the detonation products.

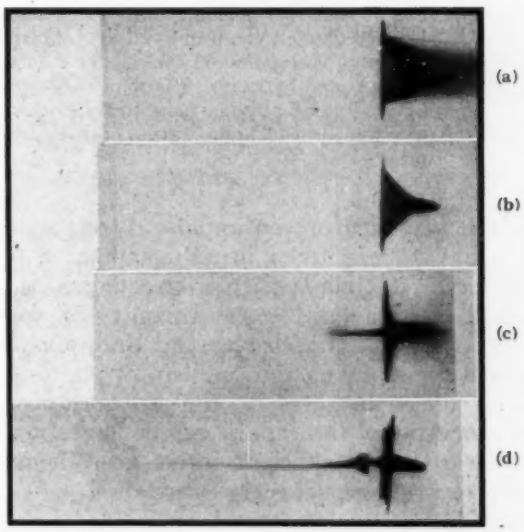


FIG. 13. Flash radiographs of metal-lined cavity charges. (a) Before detonation. The high explosive charge is cast about the 0.020-in. thick, 45° steel liner, and the charge diameter is the same as that of the base of the cone, not including the base flange. A 0.002-in. lead indicator strip is cemented along the top of the charge; (b) At approximately the instant the detonation reached the base of the conical liner; i.e., at the instant the last bit of charge detonated. The metal jet is seen inside the cone along the axis; (c) 4.8 microseconds after detonation reached base of cone; (d) 22.5 microseconds after detonation reached base of conical liner. The undisturbed base flange of the liner, the metal jet, and the slug are seen.

[14] Birkoff, MacDougall, Pugh, and Taylor, J. App. Phys. 19, 563 (1948).

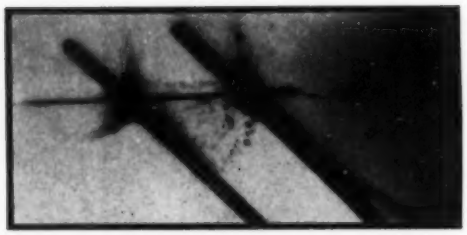


FIG. 14. Flash radiograph of a metal jet after perforating two ½-in. mild steel targets set at 45° with the direction of the jet. The slug, seen at right, moves relatively slowly and performs no part in target penetration.

Figure 14 shows a flash radiograph of one of the metal jets perforating a steel target assembly, the targets in this case being inclined at approximately 45° to the direction of the jet.

IV. CONCLUSIONS

From the examples shown in this report, it is seen that the flash radiographic technique provides a method for studying many phenomena wherein flash photography is impossible. Obviously, for phenomena involving motions exceeding a few hundred meters per second, the duration of the x-ray burst should not be greater than one microsecond, and for motions exceeding 1000 meters per second the x-ray flash should consist of a single intense burst of x-rays with duration one-tenth microsecond or less.[15] The Micronex equipment used in the work described in this report was very well adapted to cases where portability and ease of operation is required, but it did not fulfill the requirement of producing a single burst of duration not greater than one-tenth microsecond.

[15] J. C. Clark, Phys. Rev. **72**, 741 (1947).

Point-to-Plane Impulse Corona*

DONALD BAKER MOORE AND WILLIAM N. ENGLISH**
Department of Physics, University of California, Berkeley, California
(Received October 6, 1948)

The characteristics of impulse corona have been studied in a point-to-plane gap. Both positive and negative square voltage pulses of one- and two-microsecond duration were used with various voltages and pulse repetition rates. The impulse corona in air is found to be quite similar to the d.c. corona, except that the phenomena are exaggerated by the relative absence of space charge inhibition. At higher voltages, the negative impulse corona was found to have a "spike" or streamer type of discharge superimposed on the regular Trichel pulse corona. By using the nonelectron-attaching gases nitrogen and hydrogen, it is shown that these spikes are due to the negative ion space charge formed by the Trichel pulse in the gap. When the point is grounded, the intense field developed between this negative space charge and the positive space charge at the point is sufficient to cause a streamer to propagate.

THE present work on impulse corona was originally undertaken in an attempt to obtain single reproducible positive streamers. Such a method of obtaining single streamers of approximately the same amplitudes and durations would be ideal for studying the problem of photo-ionization produced in gases by the streamer advance. The properties of the impulse corona have, however, proved to be of sufficient interest in themselves that a separate report seems warranted.

Gorrill[1] has made some observations of positive impulse corona streamers in a cloud chamber which show their growth and structure. Also Hey and Zayentz[2] have made a rather complete cloud-chamber study of impulse corona using concentric cylindrical electrodes because of their interest in transmission line problems.

APPARATUS AND TECHNIQUE

The conventional hydrogen thyraton type pulser was used to give approximately square voltage

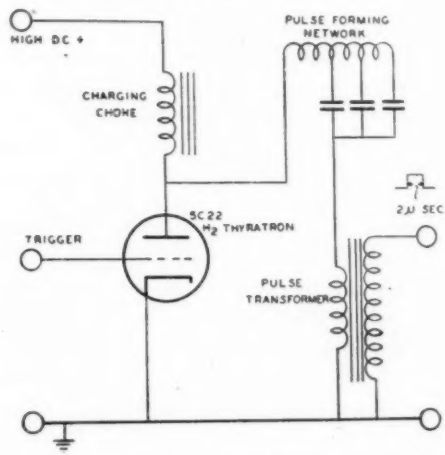


FIG. 1. Pulser circuit.

* This work was done under contract with ONR.
** Now at National Research Council, Chalk River, Canada.
[1] W. S. Gorrill, Dissertation, University of California 1939.
[2] V. Hey and S. Zayentz, J. Phys. U.S.S.R. 9, 413 (1945).

pulses of one- and two-microsecond duration at voltages up to 12,000 and a pulse repetition frequency ranging from 50 to 2000 p.p.s. Figures 1 and 2 show the circuit used and the wave shape obtained. There was no noticeable difference in the appearance of the corona for one and two microseconds, and both durations were used interchangeably. A pulse repetition frequency of 100 p.p.s. was generally used. The 12,000-volt output voltage was limited by the maximum insulation rating of the pulse transformer available.

Figure 3 shows the corona gap and the bridge circuit for separating the corona pulses from the applied high voltage pulse. The pulse transformer used here was necessary because of the ground connection in the Navy TS 28/UPN synchroscope input. The synchroscope's one-inch per microsecond sweep was triggered from the thyratron trigger unit.

The corona chamber is shown in Fig. 4 and consisted of a steel needle with a point radius of about

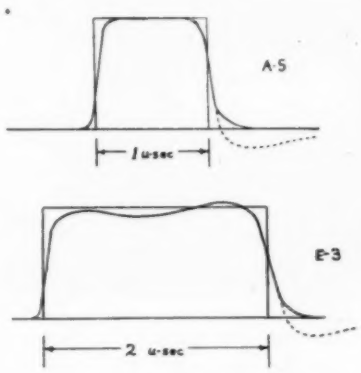


FIG. 2. Pulser output; wave form.

0.05 mm and a two-to-three-cm gap to a plane electrode which was grounded through a series resistance with a tap-off to the bridge circuit. The chamber itself was of glass, with a fine wire gauze lining at ground potential, and could be evacuated for using other gases than air at atmospheric pressure. A weak radioactive source was placed near the gap to trigger the positive corona. It had a negligible effect except near corona onset.

We have attempted to resolve the corona pulses completely from the high voltage pulse on which they are superposed, but have not been entirely successful. The space charge movements in the corona gap give a voltage pulse of a few volts in the series resistor. It was difficult to balance out the high voltage pulse of up to 12,000 volts to this accuracy because of stray capacitances and inductances, and because of small voltage fluctuations. In addition, since on the screen we view a composite picture, slight variations in the applied pulse, as

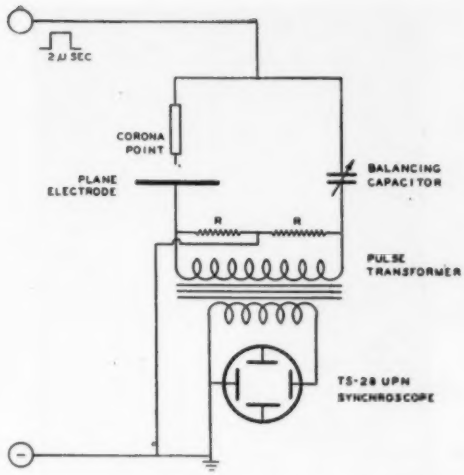


FIG. 3. Corona gap and circuit for distinguishing corona signal from high voltage pulse.

well as some a.c. pick-up with the high gain used, confuse the picture.

The synchroscope study has thus been limited to obtaining the time location of the corona processes with respect to the high voltage pulse, and to getting some idea of the intensity and nature of the corona.

A photographic study has been more fortunate. A Leica with Eastman Super XX film and a $3\frac{1}{4}$ by $4\frac{1}{4}$ plate camera with Eastman plates No. 40 and II-O were used. A five-cm Summar f:2 lens was used at an effective aperture of about f:4. Because of the nature of the corona, that is occurring only during the relatively short impulse, exposure times are measured in terms of a number of voltage pulses. Figure 5 is a typical photograph of positive and negative corona showing comparative sizes of the two discharges.

Photographs have been obtained which show the appearance of the corona discharge and the manner in which it changes with voltage. These photographs, together with visual observation of the gap and the synchroscope observations, have yielded a tentatively complete explanation of the impulse corona in terms of normal d.c. corona processes.

EXPERIMENTAL RESULTS

Positive Point in Air

In the positive impulse corona, one observes at onset a bright glow of burst pulse corona on the tip of the point, and long radial luminous filaments usually referred to as "streamers." These streamers become more intense and longer with increasing voltage, as with the positive d.c. corona streamers. They are, however, radially distributed rather than concentrated along the point axis as are the d.c. streamers. This may be seen from Fig. 6, which

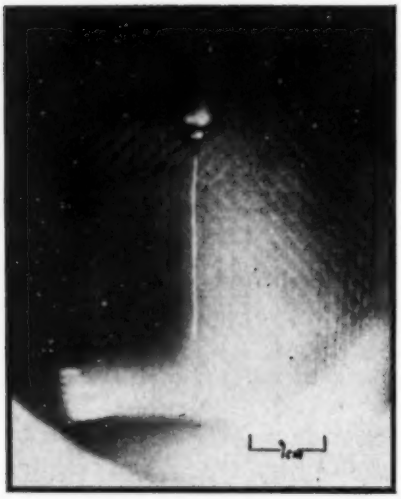


FIG. 4. Photograph of corona chamber; point radius 0.05 mm.

shows the appearance of the corona at successively higher voltages.

Attempts have been made to photograph a single streamer. We have obtained photographs of as few as ten streamers, more or less superposed, which show very clearly the filamentary nature of the streamer luminosity.

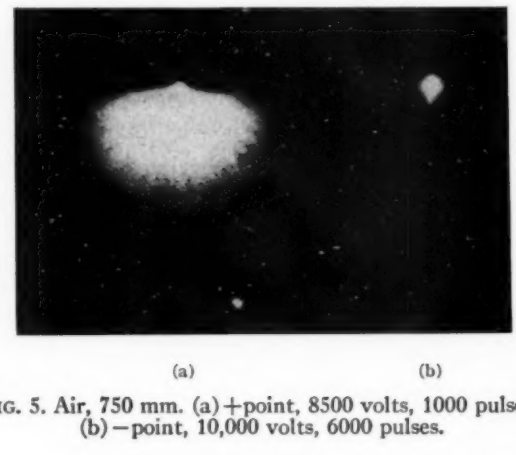
The synchroscope reveals that at low voltage a streamer may occur at any time during the impulse. As the voltage is raised, the streamer comes earlier until at about 6000 volts, where the streamers are about one cm long and very intense, the streamer pulse occurs always the same short time after the impulse (approximately 0.2 microsecond) and has a duration of about 0.5 microsecond, as reported by English[3] and Meek and Saxe.[3] There is never more than one streamer pulse during a single high voltage impulse.

Negative Point in Air

The negative impulse corona just above onset (5400 volts) differs little from the d.c. corona. On the synchroscope a Trichel pulse occurs randomly during the high voltage impulse, although not on every impulse. Visually, we observe the typical fan-shaped glow as shown in Figs. 7 and 8. As the voltage is increased to 7000 volts, the Trichel pulses become more frequent, and two or perhaps three may occur during one impulse. At about 7700 volts, the Trichel pulses begin to localize at the beginning and end of the impulse, where the voltage is slightly higher.

At about this voltage, a striking departure from the d.c. corona processes begins to develop. Visually and photographically, we observe a long narrow "spike" extending axially through the center of the Trichel pulse fan. The "spike" and the fan become more intense and extend further into the gap as the voltage is increased. At 9000 volts, two side spikes appear. The development of the "spikes" is shown in Fig. 7.

The voltage could not be raised higher because of limitations in the pulser, but somewhat the same effect can be obtained by decreasing the pressure to one-half an atmosphere. Here diffusion is greater, but does not materially change the appearance of the corona. Between about 4000 and 5000 volts, the development of the "spikes" takes place as described above. As the voltage is increased, the background luminosity in the gap out from the point increases. At about 5600 volts, the "spikes" have become broader and difficult to distinguish, and now one begins to discern luminous streaks or filaments amid the general glow. With increasing voltage these become brighter and longer, and



(a) (b)

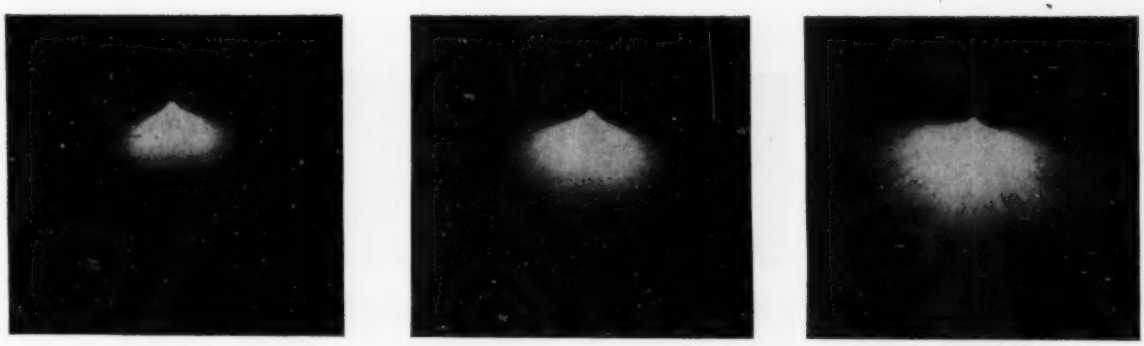
FIG. 5. Air, 750 mm. (a) +point, 8500 volts, 1000 pulses; (b) −point, 10,000 volts, 6000 pulses.

extend out beyond the general glow where they are easily seen. They appear to concentrate toward the sides rather than toward the plane, but this impression may be partially due to the much higher background luminosity in the latter direction.

At still higher voltages, approaching 9000, these "streamers" present a spectacular appearance something like a roman candle. They become many centimeters long and weave and criss-cross in the field of view. Many of them turn back parallel to the needle axis and come within a centimeter or so of the shank of the needle. At this voltage, there are secondary discharges shooting out from the shank and these, with a weaker field, often show the type of "spike" described earlier. The second picture, row 2, Fig. 8, shows the side streamers reaching out and curving back, while pictures 3 and 4 show the side "spikes" from the shank of the needle.

It is impossible to tell from observation whether the "streamers" develop out from the point or into

[3] W. N. English, Phys. Rev. **74**, 170 (1948); J. M. Meek and R. F. Saxe, pre-publication information privately communicated by the author to Professor Loeb.



(a) (b) (c)

FIG. 6. Air, 750 mm, +point. (a) 4200 volts, 6000 pulses; (b) 6000 volts, 6000 pulses; (c) 9500 volts, 500 pulses.

it, but it seems certain that they have nothing to do with the walls since the same effect is obtained in a small and in a very large chamber.

Negative Point in Hydrogen and Nitrogen

It was suggested by Dr. Loeb that the streamers seen in the negative corona might be caused by negative ion space charges in the gap and would probably not be seen in a nonelectron-attaching gas.

We, therefore, chose hydrogen and nitrogen, the two most convenient gases, to test whether the formation of negative ions was an essential part of the streamer process as observed in the negative impulse corona.

It was found that ordinary filtered tank nitrogen and hydrogen show only the regular Townsend type of glow discharge, as observed by Weissler,[4] with no trace of spikes or streamers forming in the pressure range studied and with voltages of up to 9000 volts applied to the point. This discharge is shown in column 1, Fig. 8.

The addition of one percent of oxygen to these gases did not give any evidence of streamer formation at atmospheric pressure. However, an oscilloscopic examination of the d.c. corona revealed the presence of Trichel discharge pulses characteristic of the presence of oxygen, and upon lowering the pressure to half an atmosphere and thus increasing the X/p by a factor of two, with an applied voltage of 9000, steamers were observed, as shown in column 2, Fig. 8.

Nitrogen was then studied with increasing proportions of oxygen up to that of normal air. With 5 percent added oxygen the streamers are quite clearly observed at atmospheric pressure and do not change markedly with further increase in the amount of oxygen present in the gas. Figure 8 shows the development of the negative streamer process with changes in the concentration of oxygen. It is shown at the same scale as in Fig. 4.

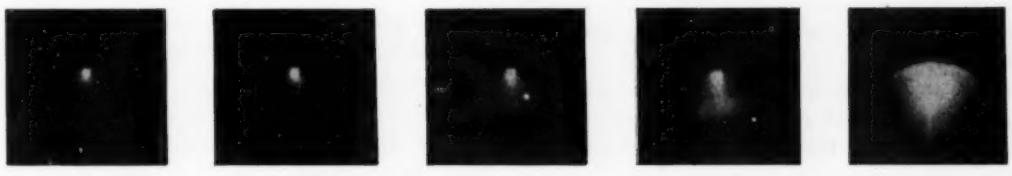
It was possible to cause the disappearance of the streamers by admitting nitrogen and pumping down to operating pressure, but because of difficulties in mixing, larger quantities of nitrogen must be admitted than would ordinarily reduce the effective oxygen concentration to lower than one percent.

DISCUSSION OF RESULTS

The primary modifying characteristic of the impulse corona discharge is the alteration of space charge distributions and the lack of over-all space charge accumulations due to ion removal between pulses except at higher voltages.

Positive Point

At onset either a burst pulse or a streamer may appear. Above this steamers are favored especially by the absence of a general positive space charge. In d.c., where the space charge is cleared most effectively only near the tip of the point, streamers radiate only from the tip along its axis. In the im-



(a) (b) (c) (d) (e)

FIG. 7. Nitrogen +5 percent oxygen at 375 mm, 100 pulses, −point. (a) 4200 volts; (b) 5600 volts; (c) 6500 volts; (d) 7800 volts; (e) 9100 volts.

[4] G. Weissler, Phys. Rev. G3, 96–107 (1943).

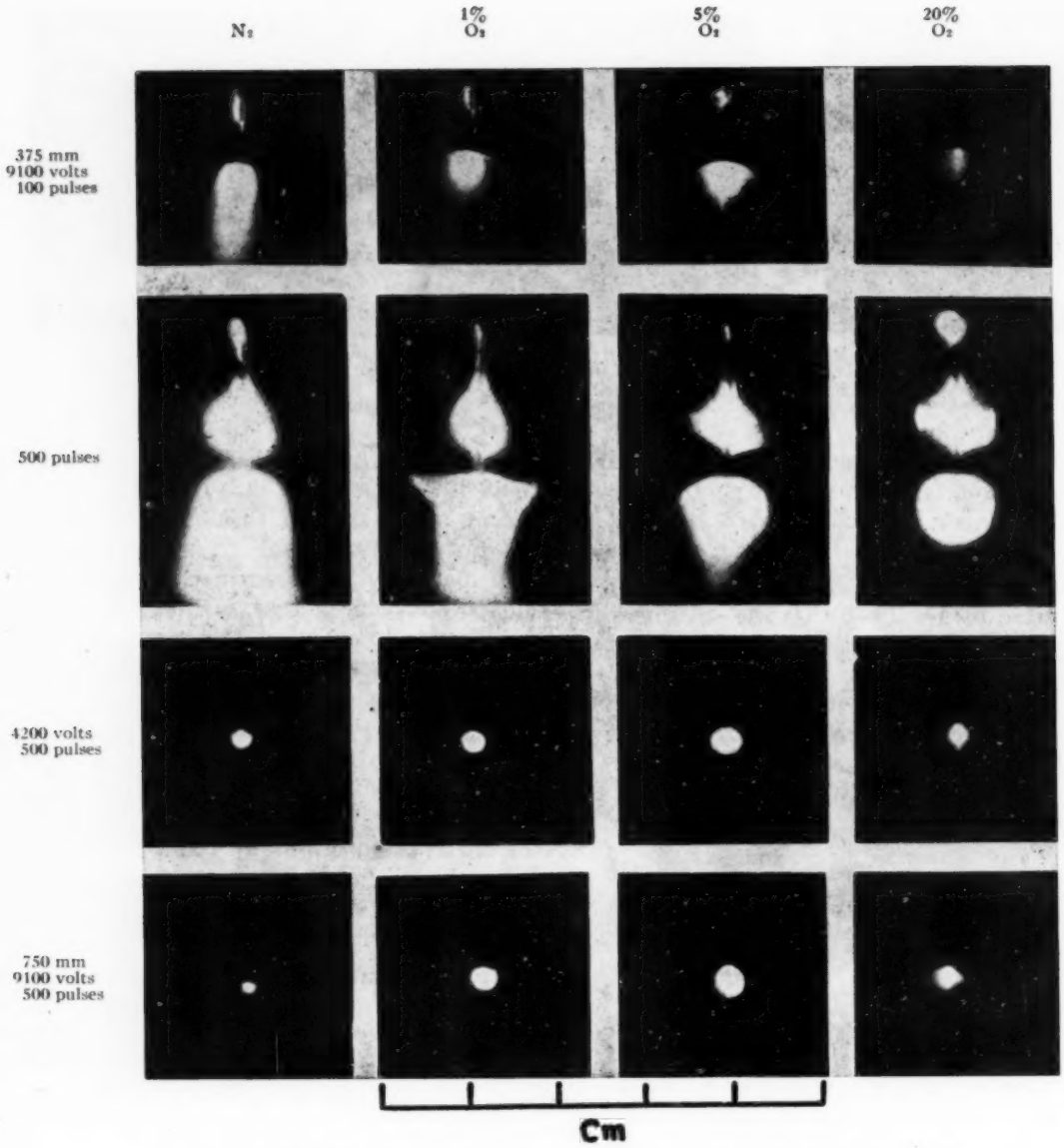


FIG. 8. Negative point corona in N_2 and O_2 mixture showing change with voltage and concentration of oxygen.

pulse corona, with the needle now relatively cleared of space charges by the intervening times between pulses and with triggering electrons approaching the needle from all directions, streamers will appear at all positions on the point and radiate along the lines of force. The streamers are relatively longer than with d.c. as they do not run into a positive space charge cloud.

At higher voltages there are sufficient negative ions and electrons left in the gap between pulses, so that a streamer can occur at the beginning of each impulse. The time necessary to clear the gap of space charge sufficiently for streamer propagation is evidently greater than the longest impulse length, two microseconds, and less than the period corresponding to the greatest p.r.f. used, 1/2000 second. This follows because there is one streamer pulse for each impulse, and decreasing the p.r.f. had no effect on the corona.

Negative Point

The negative impulse corona is quite regular in the region just above onset consisting of the conventional Trichel pulses. As the main voltage is raised, however, a most interesting situation develops. Because of the higher mobility of the negative space charge, which is electronic for at least part of its travel out of the effective choking off region, two Trichel pulses can occur on a single impulse. The second occurs at the end of the impulse and is interrupted suddenly by the grounding

of the point. This leaves a highly localized concentration of positive ions near the point, and a negative space charge of electrons and negative ions farther out in the gap. Under these conditions, Loeb has suggested that with the uncompensated high fields between positive and negative ion clouds a positive streamer type of discharge proceeds from the positive to the negative space charge.

At these potentials, the Trichel pulse discharge shows the characteristic form of a negative glow separated by a dark space from the fan-like positive glow. In this case the discharge is concentrated in the high field region at the point tip, since triggering electrons and the intense multiplication required in such pulses can occur only in high fields.

With higher potentials and larger current the space charge size is such that discharge by the central streamer spike is not adequate and radial spikes occur at the edges. The filamentary streamer processes which produce the "roman candle" effect at still higher voltages are seen to develop out of the spikes, so that it seems likely that they have essentially the same mechanism.

Extended observations of long meandering streamers at lower pressures indicate that the space charge concentrations which produce them may accumulate over many impulses and are perhaps affected by electrical wind currents in the gap. If this is so, we have a miniature "cloud-to-cloud" lightning discharge.

Hey and Zayentz have observed a very similar phenomenon in their cloud-chamber experiments with concentric cylindrical electrodes. Using a chopped impulse from a spark gap, of time duration 0.01 to 0.1 microsecond, they obtained a "rosette of short discharge paths very like positive streamers outside the inner glow."

They did not obtain this effect with 17-microsecond impulses. It seems evident that with the short chopped impulse only one discharge of the Trichel type could be initiated per impulse, and it would always be interrupted, giving the "back discharge" observed, whereas the 17-microsecond impulse was the full discharge wave from a condenser shorted to ground and thus had the relatively long decay tail which could not yield the required Trichel pulse interruption.

The observations in nitrogen and hydrogen demonstrate that the effects described unquestionably depend on the negative ion space charge for their mechanism, and cannot be ascribed to oscillations or positive overshoots on the high voltage impulse.

ACKNOWLEDGMENTS

The writers wish to thank Dr. H. J. White, Research Corporation, Bound Brook, New Jersey, for helpful discussions of pulsers and impulse corona, and for supplying circuit drawings and some key components. They are grateful to Professor L. B. Loeb for his guidance in the work and the interpretation of the results.

Thermoelectric Experiments with Extreme-Pressure Lubricants*

ROBERT SCHNURMANN
Physics Department, Manchester Oil Refinery Limited, Manchester, England
(Received October 11, 1948)

Reference is made to the conjectural nature of our present views on the mechanism of extreme-pressure lubrication. A differentiation is made between *sliding* and *tearing* so that sliding is defined as the relative motion of two solid bodies in frictional contact under the action of a tangential force with the frictional force increasing in direct proportion with the normal load. On the other hand, when the surfaces of the friction elements suffer macroscopic damage during relative motion, the frictional force increases at a more than proportionate rate with the normal load and a case of tearing is established.

The general requirement made of a lubricant is the reduction of the frictional force. The requirement made of an extreme-pressure lubricant is that it should reduce the frictional force at a sufficiently high value of the normal load for a pressure of at least 1.5×10^4 kg-wt./cm² to apply throughout the run, and in addition, the extreme-pressure lubricant should prevent tearing or scoring of the surfaces at these comparatively high bearing pressures.

A loaded revolving steel ball in frictional contact with a hardened steel block is a model of a bearing requiring extreme-pressure lubrication. The temperature elevation at the rubbing contact should be a measure of the energy dissipated by friction, and this can be determined by connecting the rubbing contact to a galvanometer which records the thermal electromotive force generated at the contact. The record of the electromotive force represents the entire frictional history of the contact in every detail. Evidence was found that extreme-pressure addition agents reduced the frictional force and extended the region of normal loads over which the frictional force was a direct proportion of the normal load.

Various types of extreme-pressure addition agents gave results which suggested the existence of an optimum dope concentration. In general, over the region of normal loads over which Amontons' law of the direct proportion between the frictional force and the normal load applied, lubricants which reduced the thermal electromotive forces generated at the rubbing contact also reduced the wear diameter. On the other hand, when tearing took place no such parallelism between friction and wear became apparent since the final wear diameter is the integrated result of the wear processes occurring during one run, whereas the record of the electromotive forces gives the details of the change of the frictional force in the course of such a run.

It is suggested that the optimum dope concentration may be a function of two counteracting effects. On the one hand, the extreme-pressure addition agent is likely to decrease the dielectric breakdown field strength of the thin layer of lubricant and to reduce thereby the electrostatic component of the frictional force. On the other hand, the lubricant becomes more corrosive with increasing concentration of the extreme-pressure addition agents. As soon as appreciable corrosion takes place, tearing sets in and causes the frictional force to increase. Thus a balance will be struck between the reduction of the frictional force by diminishing its electrostatic component, and an increase of the frictional force owing to the corrosive action of the extreme-pressure addition agents.

Evidence was found to support this latter view when metallic films were wiped out by hand on metal blocks. It was then observed that under "dry" conditions the metallic films tended to reduce the thermal electromotive forces and that the films were more effective in reducing these forces in the order tin, copper, lead. On the other hand, in the presence of an extreme-pressure lubricant of optimum dope concentration similar films *increased* the thermal electromotive force, the increase being now in the order lead, tin, and copper, in which the corrosive action might be expected to be promoted by these films.

An attempt was made to calibrate the thermal electromotive forces in terms of temperature.

INTRODUCTION

WITH the introduction of hypoid gears the limitations of ordinary gear lubrication became apparent and a long series of practical tests indicated that lubricants containing various materials, among which were certain animal and vegetable oils, sulfur, lead, and chlorine, were superior to straight mineral oils in their ability to make the gear teeth carry loads at high pressures of 1.5×10^4 to 3×10^4 kg-wt./cm² without scoring. The suggestion has been made (see, for instance, van der Minne[1] and Mougey[2]) that a film of some kind must be formed by the active constituent of the extreme-pressure lubricant reacting with the bearing surfaces, thus preventing intimate contact and consequent welding of the metal surfaces. As Simard, Russell, and Nelson[3] have pointed out,

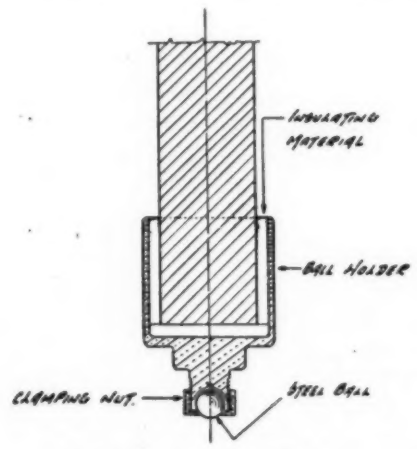


FIG. 1. Section of ball holder.

* Paper presented at the Sixth International Congress of Applied Mechanics in Paris. Section III, Paris, September 25th, 1946.

[1] J. L. van der Minne, "General discussion on lubrication and lubricants," Inst. Mech. Eng. **2**, 429 (1937).

[2] H. C. Mougey, *The Science of Petroleum*, IV, 2598 (Oxford University Press, London, 1938).

[3] G. L. Simard, H. W. Russell, and H. R. Nelson, Ind. Eng. Chem. **33**, 1352 (1941).

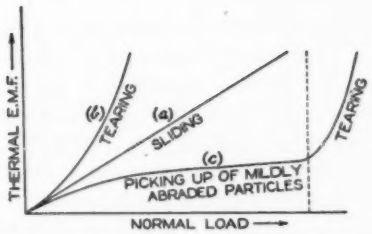


FIG. 2. Schematic representation of the increase of the thermal e.m.f. with the normal load (a) in the case of sliding, (b) in the case of tearing, (c) in the case of the steel ball picking up mildly abraded particles until eventually tearing takes place.

the diversity of extreme-pressure addition agents makes possible but few generalizations concerning the physical and chemical nature of the films which they may be supposed to form with the bearing materials.

There is general agreement that the function of any lubricant is to reduce the frictional force of a rubbing contact under given operating conditions. It is also generally appreciated that an extreme-pressure lubricant is required to reduce the frictional force when the applied normal load is high enough to give rise to pressures of at least 1.5×10^4 kg-wt./cm^2 at the end of a run when the bearing has been well worn in. An additional function of the lubricant is to prevent tearing of the surfaces of the friction elements. The problem is to consider by which physically reasonable means these qualities of a lubricant can be assessed. For the rating of extreme-pressure lubricants the petroleum industry relies hitherto on a number of testing machines, which, with the exception of the four-ball apparatus,[4] do not measure a definite physical quantity. It is clear, however, that any mechanism which allows one to record either the frictional force itself or a function of it in dependence upon the normal load directly rates lubricants for a given pair of friction elements subjected to a given speed of relative motion. The order of merit is a decrease of the frictional force for given values of the normal load which are permissible at a given speed before appreciable tearing of the solid surfaces takes place. Whenever a tangential force is made to act at the common boundary of two solid bodies, relative motion can take place either by sliding or by tearing. In the former case the surfaces remain undamaged and Amontons' law applies;[5,6] this ex-

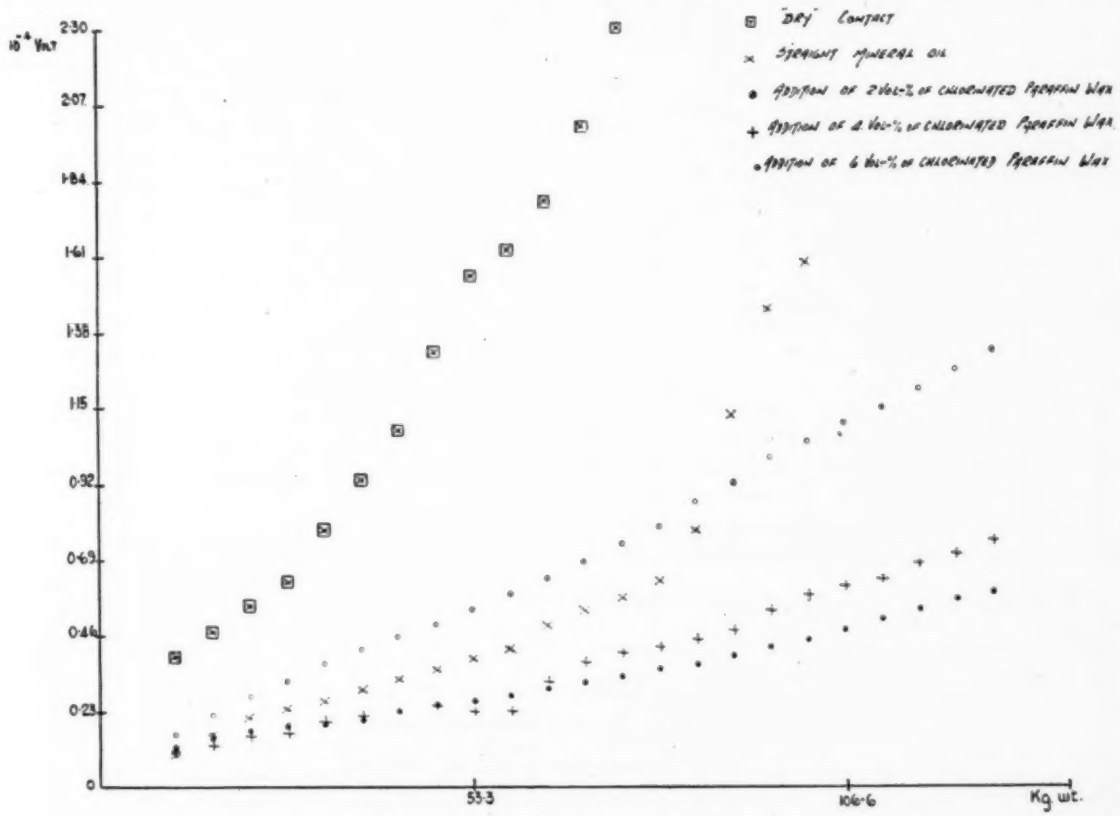


FIG. 3. Electromotive force-load diagram of steel balls revolving (18 rev./sec.) on a Timken test block. The normal load was increased by 5.3 kg-wt. every 10 seconds.

[4] D. Boerlage, Engineering **136**, 46 (1933). R. Schnurmann, Engineering **149**, 567 (1940); **150**, 236 (1940).
[5] Sir W. B. Hardy and I. Doubleday, Proc. Roy. Soc. **A100**, 550 (1922).
[6] R. Schnurmann, Rep. on Progress in Physics **8**, 71 (1941).

presses the frictional force as a *constant* proportion of the normal load for a given combination of friction elements. On the other hand, when tearing takes place, i.e., when the surfaces suffer macroscopic damage, the frictional force increases at a *more than proportionate* rate with the normal load.

A loaded revolving steel ball in frictional contact with a hardened steel block is a model of a bearing requiring extreme-pressure lubrication. Frictional heat is generated at the contact area, and the rate of heat generation depends upon the magnitude of the frictional force at a given rate of revolution of the steel ball. Thus, it is easy to find that without a lubricant the steel ball becomes hot to the touch after a few minutes running when the normal load has been increased in steps of about 5 kg-wt. to about 150 kg-wt. On dismantling this frictional contact, a comparatively large wear impression is found both on the steel ball and on the flat surface, and this impression shows signs of bad tearing. On the other hand, the application of a small drop of an extreme-pressure lubricant, such as a mineral oil containing two percent of chlorinated paraffin wax, for instance, keeps the bearing cool under similar running conditions, and on dismantling this frictional contact the wear impression is found to be very small and quite smooth.

The temperature elevation at the rubbing contact should be a measure of the energy dissipated by friction, and this can be recorded by connecting the rubbing contact to a galvanometer which indicates the thermal electromotive force generated at the contact. It has been shown before[7–9] that the two metal elements of a frictional contact can be

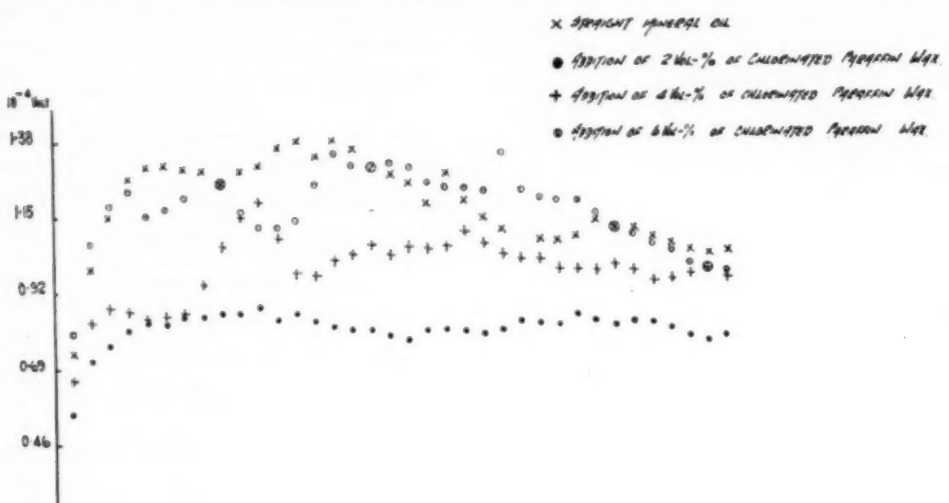


FIG. 4a. The electromotive force at constant normal load (40 kg-wt.). Steel balls revolving (18 rev./sec.) on the flat surface of a steel roller.

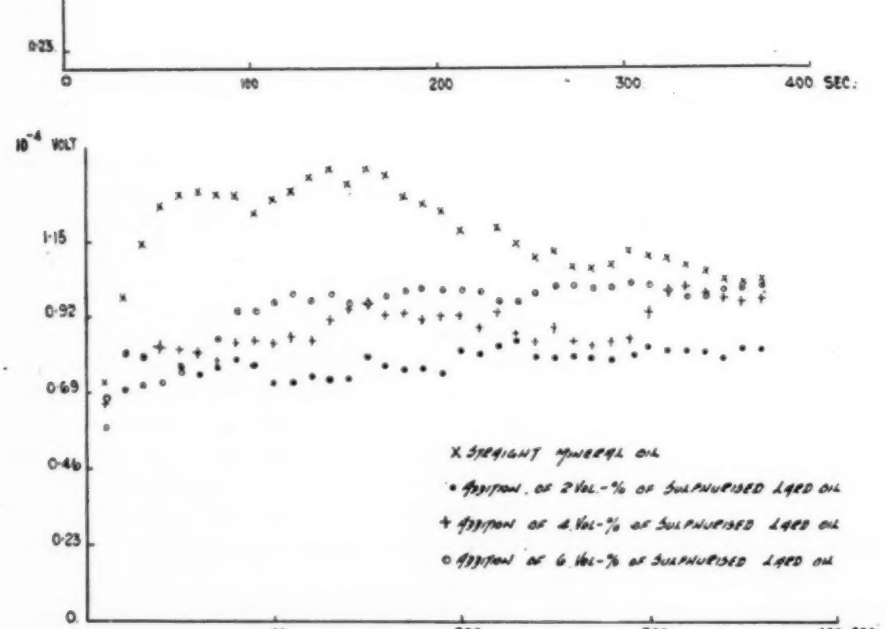


FIG. 4b. The electromotive force at constant normal load (40 kg-wt.). Steel balls revolving (18 rev./sec.) on the flat surface of a steel roller.

[7] E. G. Herbert, Proc. Inst. Mech. Eng. 1, 289 (1926).
[8] F. P. Bowden and K. E. W. Ridler, Proc. Roy. Soc. **A154**, 640 (1936).
[9] F. Morgan, M. Muskat, and D. W. Reed, J. App. Phys. 12, 743 (1941).

made the two elements of a thermocouple by connecting them to a galvanometer which records the temperature at the *actual* area of contact. The recorded temperature depends at every instant upon the equilibrium between the rate of heat generation and the rate of heat dissipation, the former being in direct proportion to the speed of relative motion, and the force of friction, and the latter depending upon the thermal diffusivities of the friction elements, their geometrical dimensions, and the heat dissipation from their surfaces by convection and radiation.

EXPERIMENTAL METHOD

The experimental set-up was very simple; it required a rigid clamping device for a steel ball, so that a normal load could be applied when the steel ball was made to revolve on the flat surface of a clamped piece of metal. An electrically driven drilling machine incorporates all the required movements and needs only slight adaptation such as the making of the ball holder and the fixing of a loading lever to the pinion axle. It was essential that the water equivalent of the ball holder should be small. A hemispherical cup was cut *in situ* into the bottom of the ball holder (Fig. 1) to ensure true running of the steel ball which was prevented by a clamping nut from turning in the carefully ground cup during an experiment. $\frac{3}{8}$" diameter steel balls (commercial ball bearings) were used throughout these experiments. The circuit for measuring the electromotive forces incorporated a brush contact. The voltages were calibrated in terms of temperature with the steel ball at rest and loaded with 122.6 kg-wt. against the flat surface of a clamped piece of metal, and with the "hot" junction of a thermocouple clamped to the particular piece of metal. The clamping device for the metal block consisted of a large amount of metal through which heat was conducted to the contact area of the steel ball in a way which should ensure that the temperature recorded by the thermocouple was the same as that at the contact area.

Two modes of operation of the one-ball apparatus can be applied. Either the thermal electromotive force is measured as a function of the normal load at a given rate of revolution of the steel ball, or else the thermal electromotive force can be recorded as a function of the time of running at a given normal load and at a given rate of revolution of the steel ball. In the former case three types of curves can be obtained depending on whether sliding or tearing takes place or whether the steel ball picks up some mildly abraded particles (Fig. 2). In the case of *sliding*, the thermal electromotive force increases in direct proportion with the normal load. In the case of *tearing* the thermal electromotive force increases at a *more than proportionate* rate

TABLE I. Diameters of final wear-impressions in mm.

Wiped out metal film	"Dry"	Straight mineral oil	Vol.-percent of chlorinated paraffin wax in mineral oil 2	4	6	8	10	20	
No film	2.12	1.61	0.79	1.58	1.48	1.68	1.37	0.94	
Pb	1.29	1.53	0.77	0.78	0.76	—	0.80	0.87	Timken
Sn	1.52	1.35	0.80	0.80	0.78	—	0.98	1.00	test
Cu	1.36	1.48	0.78	1.37	1.10	—	—	—	block
Zn	1.64	1.38	0.80	1.54	1.40	—	0.71	0.88	
No film	1.40	1.54	0.93	0.75	1.45	0.79	0.71	0.98	Skefco
Pb	1.30	1.16	0.71	0.78	1.15	1.07	0.97	1.03	steel
Sn	1.21	0.80	0.72	1.42	1.42	0.80	—	—	roller
No film	2.82	1.94	1.83	1.96	1.94	—	1.88	—	Copper
Pb	3.00	1.91	1.79	1.89	1.96	—	1.95	—	block
Tinned*	3.24	2.00	2.43	2.71	2.78	—	2.98	—	
No film	4.07	—	1.46	1.40	—	—	—	—	Brass block

			Vol.-percent of lead oleate in mineral oil 2	4	6	
No film	2.12	1.61	1.00	0.92	1.03	Timken test block
No film	1.40	1.54	0.78	1.07	1.21	Skefco steel roller

* This coating had not been obtained by wiping out tin but by tinning properly one face of the copper block.

with the normal load. When tearing took place and appreciable electromotive forces had been generated, the observation was made that on arresting the steel ball the galvanometer deflection continued to increase for a little while, owing to the decreased convection from the ball and from the ball holder, before falling off exponentially. In the third instance, when the steel ball picks up mildly abraded metal particles, the thermal electromotive force increases at a less than proportionate rate with the normal load, until eventually at a sufficiently high normal load tearing takes place. The case of picking up of mildly abraded particles can be directly verified by experiments with blocks of either copper or brass when the contact area of the steel balls acquires a coating of the respective metal, which is visible with the naked eye.

With the other procedure where the thermal electromotive force is recorded as a function of the time of running at given values of the normal load and of the rate of revolution of the steel ball, the thermal electromotive force climbs rapidly to an approximately constant value which is reached

TABLE II. Friction and wear changes owing to wiped out metallic films.

Metal block	Wiped out film	Percentage change of electromotive force (in the approximately linear region of the electromotive force-load curve)	Percentage change of diameter of final wear impression
Timken test block	Tin	−18.5	− 9
	Copper	−37.5	−18.6
	Lead	−50	−22.8
Hardened steel roller	Tin		+21
Copper block	Lead		+ 6.4
	Tin		+14.8

when equilibrium has been obtained between the rate of generation of frictional heat at the contact area and dissipation of this heat from the contact area by diffusivity through the friction elements and by convection and radiation from their surfaces. For given operating conditions the height of the approximately constant level to which the electromotive force rises depends on the magnitude of the frictional force.

RESULTS

"Dry" Contacts

The electromotive force was observed to increase with the speed of relative motion when the steel balls were rotated at the rate of $8\frac{1}{3}$, 18, 34, and 72 rev./sec., respectively. The revolving steel balls generate their own wear impression so that in the course of one of these experiments the peripheral speed at the contact area increases although the rate of revolution of the steel ball is maintained constant. The peripheral speed was of the order of a few centimeters per second (as calculated from the number of revolutions per second and the diameter of the *final* wear impression), and the pressure was of the order of 1.5×10^4 kg-wt./cm^2 (as calculated from the applied normal load, the diameter of the steel ball, and the diameter of the *final* wear impression).

The spinning steel balls were found to generate appreciable electromotive forces even in contact with steel, in which case an appreciable amount of red iron oxide was abraded under "dry" conditions. The direction of the thermoelectric current was from the contact of the steel ball with the flat piece of metal to the brush contact for steel rollers, Timken test blocks, mild steel strips, copper blocks, brass blocks, and Nichrome strips, whereas it was in the opposite direction for stainless steel strips, gudgeon pins, and phosphorbronze strips.

The above-mentioned tendency of the steel balls to pick up metal from the flat surfaces became clearly apparent after they had been revolving on copper, brass, and phosphorbronze, respectively; their contact area had then acquired a coating of these metals. The galvanometer records indicated when the picking up of metal took place by an appreciable decrease in the rate of change of the electromotive force with increasing load. However, even in cases of pronounced picking up of metals, the electromotive force begins to increase at a more than proportionate rate at sufficiently high values of the normal load for the surfaces to become torn.

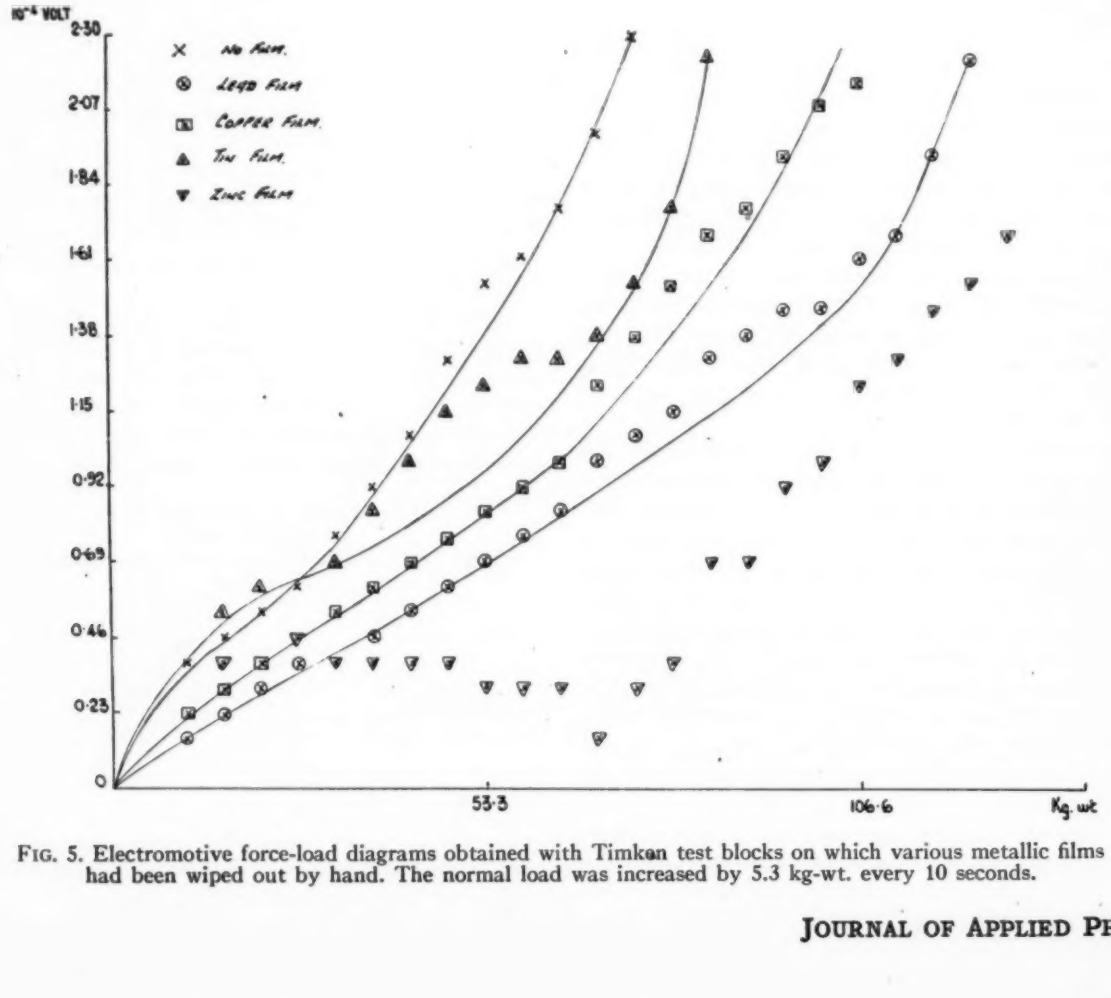


FIG. 5. Electromotive force-load diagrams obtained with Timken test blocks on which various metallic films had been wiped out by hand. The normal load was increased by 5.3 kg-wt. every 10 seconds.

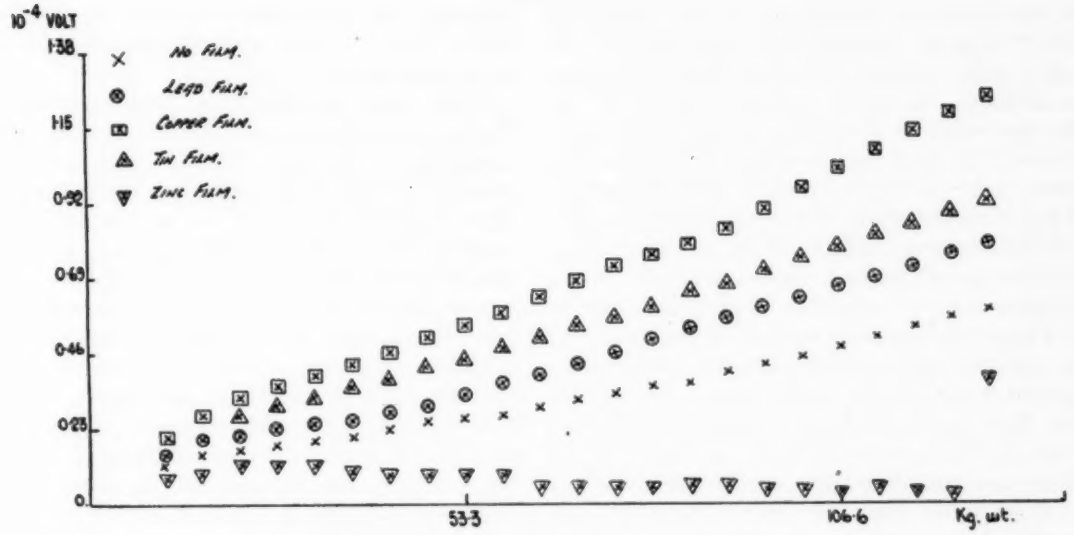


FIG. 6. Electromotive force-load diagrams obtained with Timken test blocks on which various metallic films had been wiped out by hand. A drop of mineral oil containing 2 vol.-percent of chlorinated paraffin wax was applied in each case. The normal load was increased by 5.3 kg-wt. every 10 seconds.

The electromotive force was larger when hard chips were worn off the flat surfaces of Nichrome, for instance, than when the steel balls picked up some soft metal or abraded a fine dust.

Lubricated Contacts

For similar operating conditions the electromotive forces were found to be smaller in the order: no oil, white oil, straight mineral oil. Extreme-pressure addition agents applied to a mineral oil as a rule further reduced the electromotive force and extended towards higher normal loads the region over which Amontons' law was obeyed (Fig. 3). The measurements, both by the method of recording the thermal electromotive force as a function of the normal load and of recording it as a function of the time of running at a given normal load, exhibited an optimum concentration of the extreme-pressure additive (Fig. 4).

From a large number of experiments with various types of extreme-pressure addition agents such as chlorinated paraffin wax, other chlorinated materials, sulfurized lard oil, zinc naphthenate, and lead oleate, the observation followed that the final wear impression was conspicuously small and smooth at the optimum dope concentration of each of the addition agents. It was found that the diameters of the final wear impressions on Timken blocks and on hardened steel rollers, for instance, were smallest when a few percent of an extreme-pressure addition agent, in a number of cases as little as two percent, had been added to a straight mineral oil and when the record of the electromotive force indicated no transition to tearing (Table I).

When tearing takes place, deviations from Amontons' law arise, and under stationary operating conditions the electromotive force must then increase at a more than proportionate rate since it is a function of the frictional force. Therefore, tearing proceeding for some time must lead, on the one hand, to a considerable electromotive force and, on the other hand, to a large diameter of the final wear impression. However, the processes taking place at the area of contact may change in the course of one run, and while the record of the electromotive forces reveals all the details such as picking up of metal, wearing off of fine particles, and tearing, the diameter of the final wear impression gives an integrated result only, so that in these cases no relation between the force of friction and the rate of wear can be established. This point is illustrated by the results summarized in Table II, where a parallelism is indicated between the percentage change of the electromotive force and of the diameter of the final wear impression when films of tin, copper, and lead, respectively, have been wiped out on Timken test blocks. On the other hand, films of tin, and of lead and tin, respectively, wiped out on a hardened steel roller and on a copper block, respectively, gave rise to tearing and to an increase of the final wear impression, so that in these three cases no relation could be established between the percentage change of the electromotive forces and of the final wear diameters.

Wiped Out Metallic Films

From other experiments[10] it would appear that particularly under conditions of slow relative mo-

[10] R. Schnurmann and E. Warlow-Davies, Proc. Phys. Soc. **54**, 14 (1942).

tion the electrostatic component of the frictional force may assume appreciable proportions. It would thus seem possible that one function of an extreme-pressure addition agent would be to diminish the value of the electrostatic component of the force of friction by decreasing the dielectric breakdown field strength of the thin boundary layer. This decrease could not be expected to continue indefinitely, particularly as the extreme-pressure addition agents exert a corrosive effect on the metal bearing surfaces and thus tend to give rise to tearing. These two opposed actions of the extreme-pressure addition agent might balance each other at a concentration which would appear to be the optimum dope concentration. Indirect confirmation of this would seem to derive from observations with wiped out metallic films which could be expected to promote the corrosive action of the extreme-pressure agent.

Metallic films wiped out on metal surfaces by hand were found to affect both the magnitude of the electromotive force and the diameter of the final wear impression. When no oil had been applied to the rubbing contact, films of tin, copper, and lead, respectively, wiped out on Timken test blocks reduced the electromotive force in this order of merit (Fig. 5), and also the diameter of the final wear impression.

With some metallic films wiped out by hand on Timken test blocks, the electromotive forces were appreciably smaller when a drop of a straight mineral oil had been applied to the rubbing contact than without the oil. Whereas the addition to the mineral oil of two percent of chlorinated paraffin wax,[11] for instance, still further reduced the electromotive forces, the effect of the metallic films was now to increase the thermal electromotive forces (Fig. 6), tin giving a larger increase than lead, and copper giving a still larger increase than tin. With a still higher dope concentration, with an addition to the oil of four percent of chlorinated paraffin wax, for instance, the film of copper gave rise to pronounced tearing.

The temperature calibration confirms the conclusion drawn from the recording of the electromotive forces that in the case of an addition of two percent of chlorinated paraffin wax to the oil the metallic films appreciably increase the frictional force, while the discrimination between the various metallic films is less pronounced on the

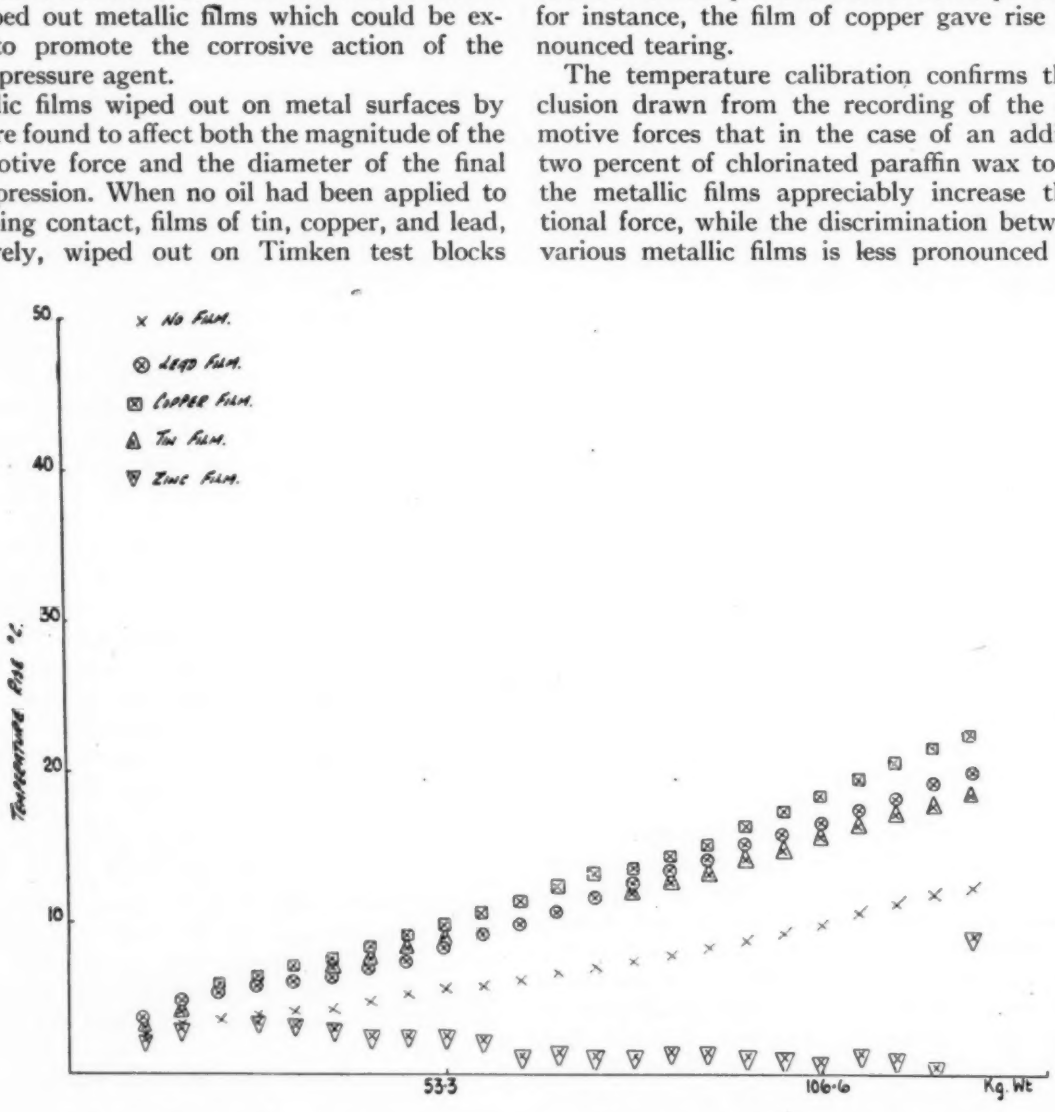


FIG. 7. Replotting of the data recorded in Fig. 6 making use of the temperature calibration of the electromotive forces (Table III).

[11] The chlorinated paraffin wax used for these experiments contained 42 percent of chlorine by weight. Its density and viscosity were at 25°C: 1.160 g/cm³ and 25 poises, respectively.

temperature scale than on the scale of the thermal electromotive forces (Fig. 7).

Remaining to be considered is the question as to how the thermoelectric force between friction elements covered with various metallic films can be regarded as a measure of the frictional force. Calibration data of the thermoelectric forces in terms of temperature obtained under static conditions and at a normal load of 122.6 kg-wt. show that the effect upon the thermoelectric force of the metallic films wiped out by hand on Timken test blocks is fairly small (Table III), requiring a temperature rise by about 20°C for each 10^{-4} volt of the thermal electromotive force. With blocks of copper or brass, respectively, a temperature rise by about 62°C would be required for each 10^{-4} volt, and in the case of a lead film wiped out on a brass block, a still higher temperature rise of about 97°C would generate a thermal electromotive force of 10^{-4} volt.

TABLE III. Thermoelectric forces between a clamped ½" diameter steel ball and various metal blocks under static conditions at high load.

Steel ball loaded against	Temperature rise in °C for each 10^{-4} volt
Timken test block	20.8
Copper film on Timken block	18.2
Tin film on Timken block	20.0
Lead film on Timken block	25.2
Zinc film on Timken block	29.5
Copper block	62.6
Brass block	62.6
Lead film on brass block	97.3

DISCUSSION

The method described here of determining and comparing extreme-pressure properties of oils involves, in common with Boerlage's four-ball apparatus, the employment of ball-bearing steel balls which are universally obtainable and highly standardized. This, together with the fact that one steel ball suffices for determining a complete electromotive force-load curve at a given rate of revolution of the steel ball, and with the possibility of making numerous wear impressions on one surface of the flat metal block, ensures a high degree of reproducibility. The amount of oil required for each experiment is sufficiently small to introduce no appreciable thermal diffusivity. Furthermore, the use of a small drop of oil allows of the observation of the dirtiness and possible frothing of the oil drop on dismantling the friction elements at the end of each experiment, since there is no inrush of oil to the contact area.

The observation was made that the quality of the lubricant affected the galvanometer readings even in cases of severe tearing, when hard chips were worn off the flat surface of nichrome, for instance, and this observation would appear to support the view that the extreme-pressure additions reduced the frictional force rather than contributing to the formation of films of increased "rupture strength."

Modification of Silica Replica Technique for Study of Biological Membranes and Application of Rotary Condensation* in Electron Microscopy

F. HEINMETS**
*Research and Development Department, Camp Detrick, Maryland****
(Received October 11, 1948)

A modified silica replica technique for electron microscopy has been developed involving direct condensation of silica upon the object. For improvement of the silica replica technique in general, the following new procedures are introduced:

(a) The undesirable effects of brittleness of silica replica is largely obviated by use of a thin plastic supporting film.

(b) A working method is introduced which completely avoids manual manipulation of the unsupported replicas.

(c) A better definition of small structural elements of the specimen is obtained when a new method of condensing metallic vapors on a rotating replica is introduced; when compared to the usual shadowcasting, an improved two-dimensional definition of the objects is achieved; also the combined use of rotary condensation and shadowcasting provides a better general definition of objects than would be obtainable by either method alone.

Reproducibility of the morphology of crystals and particles by the modified silica replica is demonstrated with the electron micrographs.

INTRODUCTION

THE polystyrene-silica surface replica technique has been successfully applied by Heidenreich, Peck, Gerauld, and their associates[1–3] to studies of metallic and various other surfaces. Our attempts to apply those techniques for study of some structural interfaces of biological systems have not been successful. The following paper will describe the development of a modified silica replica technique and its application to the study of the dried chorio-allantoic membrane of the chick, a widely used tissue for the growth of many animal viruses.

DEVELOPMENT OF A MODIFIED SILICA REPLICA TECHNIQUE

The first attempts to obtain replicas from dried chorio-allantoic membranes were made using a liquid-plastic and silica technique.[4] Collodion or Formvar was employed as the liquid plastic. There were considerable difficulties in separating the dried plastic layer from the membrane surface. Microscopic observations revealed also that little of the replica surface was entirely free from tissue fragments, and only a few areas were large enough for photographing. Furthermore, inspection of photomicrographs often showed attachment of a large number of small fragments to the replica membrane surface. As the purpose of the experiments was a systematic study of the entire membrane surface, the method just described was considered unsatisfactory. A solution to the problem was found when silica was condensed directly on the surface of the membrane.

There are a number of difficulties associated with the silica replica technique in general. As a result of the fragility of silica films one has to work with low electron beam intensity, making visual observation and focusing difficult. Manipulating the thin silica film in a plastic solvent ("fishing") is very arduous, time consuming, and often results in the loss of valuable preparation. When the vertical gradient of the structural elements is small, the silica replicas offer rather poor contrast.

In order to reduce the previously mentioned difficulties in developing a modified silica replica technique, various new procedures were introduced. The undesirable effects of brittleness of silica replica were largely obviated by use of a thin plastic supporting film. In order to avoid unsatisfactory "fishing" techniques, a working method was introduced which completely avoids manual ma-

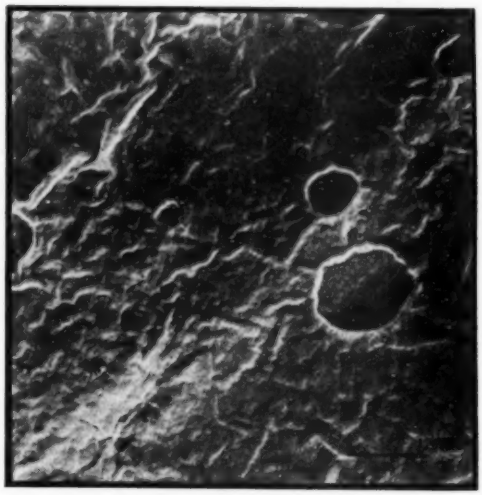


FIG. 1. Silica replica of chorio-allantoic membrane; gold shadowcasting. Negative print.

* Rotary condensation is used as abbreviation for process of condensing vapors of solids on a rotating electronmicroscope specimen.

** With the Technical Assistance of J. T. Quynn.

*** Present address: Naval Medical Field Research Lab., Camp Lejeune, North Carolina.

[1] R. D. Heidenreich and V. G. Peck, J. App. Phys. **14**, 23 (1943).

[2] C. S. Heidenreich, C. H. Gerauld, and R. E. McNuety, Tech. Pub. 1946, Metals Tech. 13, 1979.

[3] C. H. Gerauld, J. Dent. Res. **23**, 239 (1944).

[4] C. H. Gerauld, J. App. Phys. **18**, 333 (1947).

nipulation of the unsupported replicas, all necessary handling being performed while the film is supported by a heavy plastic layer. A better definition of small structural elements of the specimen was obtained when a new method of condensing metallic vapors on a rotating replica was introduced. This method is performed essentially the same way as the usual shadowcasting[5] method, except that the specimen is rotated while the shadowcasting metal is evaporating from the filament. With this method, in electron micrographs, there is practically no visible shadow around the particle or profile, but the density of the particle or surface profile is considerably increased. This subject will be discussed more completely in another section of the paper.

The final replica preparation procedure developed is as follows: The chorio-allantoic mem-

FIG. 2. Silica replica of chorio-allantoic membrane; gold shadowcasting combined with rotary condensation. Negative print.

brane is removed from the chick embryo and thoroughly washed, first in saline and then in distilled water. Pieces a few square centimeters in area are dried on microscope slides and again rinsed a few times in distilled water. The membranes are placed in the vacuum chamber and a layer of silica is condensed at normal incidence on the surface in the usual manner. The distance between the membrane and the spiral tungsten wire cone, used as a container for the evaporation of powdered quartz, should be large enough to avoid overheating of the tissue. Experience has shown that 20-cm distance is sufficient; the amount of silica used varied between 5 and 15 mg, depending on the nature of the replica. Subsequently the membranes are shadowcast, thus producing a positive surface replica. An alternative

[5] R. C. Williams and R. W. C. Wyckoff, J. App. Phys. **15**, 712 (1944).

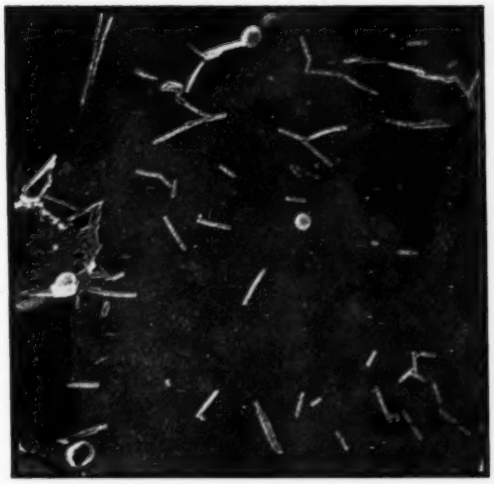


FIG. 3. Silica replica of tobacco mosaic virus; rotary condensation only. Negative print.

method has been used; first, a very thin layer of quartz is condensed on the membrane surface; then shadowcasting is performed and a thicker layer of quartz is deposited upon the shadowcast specimen. The purpose of the second layer is to give sufficient strength to the replica and at the same time provide a protective coating for the metal which might otherwise be displaced during the later phases of replica preparation, especially while the supporting plastic layer is being dissolved. The final thickness of the combined quartz layers has to be selected according to the nature of the surface: Fine structure studies require thin films, and deeply profiled surfaces require strong heavy films.

In most cases usual shadowcasting of the replicas is sufficient, but in some cases rotary condensation combined with usual shadowcasting is preferable because it produces a replica which appears to give a better defined surface structure. The next step is to apply to the membrane surface and nearby areas a series of layers of collodion. The collodion solutions used are 0.5–4.0 percent in amyl acetate, one of the more dilute solutions being recommended for the first layer in order to get better binding on the surface. Subsequent layers can be heavier, and the final one should be heavy and strong enough to permit manipulation by hand. To the edge of the dry collodion layer is attached a piece of Scotch Tape and the entire layer is pulled off very slowly by applying force at a small angle. When the collodion layer is almost completely free, another slide, wetted with a few drops of water, is placed below the free part of the film and the latter is transferred to it, the negative side of the replica facing upwards. Then the lower end of the collodion film is freed from the first slide and the Scotch Tape cut off. The collodion film sometimes becomes wrinkled during this process, but can be smoothed out by

careful manipulation, and then allowed to dry. Subsequently the slide is immersed in a 0.05–0.10 percent solution of Formvar in chloroform, leaving a thin supporting Formvar film on the silica. We have found that this thin Formvar film does not noticeably decrease the definition of the replica, even when such small objects as tobacco mosaic virus are being studied. On the other hand, it adds so much to the mechanical strength of the replica that intermediate and sometimes strong beam intensities can be used for systematic visual observation. After application of the Formvar layer, the replica is transferred to another wet slide, so that the Formvar layer is in contact with the glass surface. One end of the film is then lifted from the wet slide and several of the usual specimen-supporting screens, with edges flattened, are placed in the desired positions on the slide and the film placed over them. Upon drying, the film spreads smoothly over the specimen screens. The slide is then placed in a shallow vessel and enough amyl acetate added to cover it. When collodion is dissolved, the amyl acetate is removed from the dish by very careful pipetting, since there is the danger that the specimen screen beneath the thin replica film may move and break the replica. Then the slide is dried again. The screens, at this point, are covered with a silica replica supported by a thin Formvar film and are ready for use. Observations show that the final replica is dried on the specimen screen in a smooth layer, and for less precise work, shadowcasting of the replica can be performed directly on the screen. It should be emphasized that in our experience, satisfactory electron micrographs of replicas can be obtained only when an objective aperture is used in the microscope. Otherwise, contrast is so poor that small structural detail is lost and even larger surface profiles are ill-defined. All pictures presented here were taken with an objective aperture, using the R. C. A. type A microscope.

TECHNIQUE OF ROTARY CONDENSATION

Rotary condensation has been performed in the RCA EMV-1 vacuum unit. A small electric motor is mounted on the base of the vacuum chamber, and a current is supplied through the terminals provided for an outside power source. At the upper end of the shaft is attached a disk which was provided with phosphor-bronze clips for holding specimen screens or replicas. When shadowcasting more than one specimen screen at a time, exact center placement of all the screens is not possible, so that differences in angles of incidence and distance from the vapor source are introduced, but these are of little significance when the source is at a distance of 20–30 cm. When the correct vacuum in the chamber has been reached, the platform is rotated at a speed of about 2 or 3 rotations per second. Then a slow evaporation process, as in the usual stationary shadowcasting method, is commenced. The amount of shadowcasting metal to be used depends upon the type of specimen and the character of the information desired. For example, for a heavy rotary shadowcasting of a specimen on a collodion supporting film, 4 cm of 10-mil gold wire at a source distance of 13 cm and height of 1.5 cm was used. The presence of a motor with conventional windings in the vacuum chamber increased to three hours the required pumping time for about 0.2 micron of mercury. For a more practical set-up, a different type of motor or an outside magnetic drive could be used.

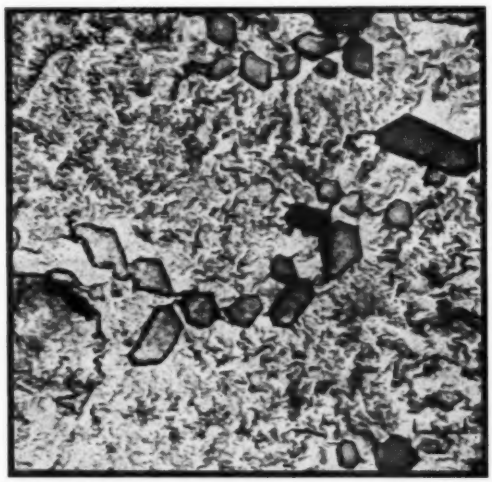


FIG. 4. Silica replica of salt crystals on chorio-allantoic membrane; without shadowcasting. Positive print.

REPLICAS OF BIOLOGICAL MEMBRANES***

Figure 1 shows an electron micrograph of positive replica prepared from the chorio-allantoic membrane of a nine-day chick embryo by the modified silica replica technique with usual gold shadowcasting only. A few large cavities and general surface irregularities are distinctly visible. Using shadows for orientation, there is no difficulty in visualizing the elevations and depressions. Figure 2 also reproduces a micrograph of a positive replica of the membrane; here a combination of usual shadowcasting with rotary condensation was used. Finer surface structure is apparently resolved. In order to test the reliability of the modified replica technique in giving faithful representation of rod-shaped particles, a replica was taken from tobacco mosaic virus which had been dried on a formvar-covered glass slide, using a relatively thin silica film and rotary condensing alone. The rod-like features in the replica (Fig. 3) are well-defined and resemble

*** The length of the scale on each of the electron micrographs is one micron.

the original particles. There are also a few round particles with less dense central areas, indicating that this region had unfavorable angles for building up density by condensing. A second evaporating source at a larger angle could be used for improving the contrast in such a central area. The large dark areas in the picture represent regions of broken replica layers.

REPLICAS FROM SALT-CRYSTALS

Questions arise as to the reproducibility of surface profiles and other geometrical forms and the degree of distortion introduced by the modified replica technique. For information on this point, a set of replicas were prepared from various inorganic crystals and results indicated no visible deformation of crystal shapes. Figure 4 shows a replica prepared without shadowcasting from a chorio-allantoic membrane on which crystals have formed. The profiles are mostly defined by straight lines, indicating that manipulation of the replicas has not produced two-dimensional distortion. Reproduction of the crystal form is unsatisfactory, however, because of the absence of relief. Figure 5

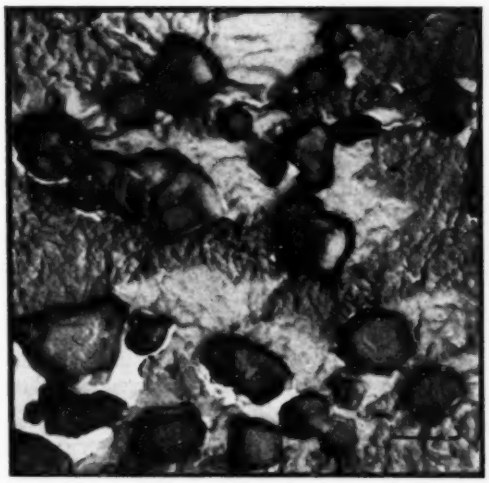


FIG. 5. Silica replica of salt crystals on chorio-allantoic membrane; gold shadowcasting combined with rotary condensation. Positive print.

shows a micrograph of a replica made from a similar object to that used for Fig. 4, but here usual shadowcasting has been combined with rotary condensation. Various crystal shapes have been reproduced and a three-dimensional effect achieved. Light areas adjacent to the crystal replicas are shadows formed by the shadowcasting. The density of crystal replica is mainly increased by rotary condensation. A further aspect of the use of the usual shadowcasting combined with rotary condensation is demonstrated by Fig. 6. The replica represented here was prepared from potassium chloride particles on Formvar covered glass. By using rotary condensation, the electron-optical density of the silica shell of the particle is increased to such a degree that it resembles the original particle, while the superimposed shadowcasting provides relief. In final appearance, the electron-micrograph of this replica resembles one prepared directly from the same specimen by shadowcasting.

DISCUSSION

An attempt will be made to analyze what improvements the rotary condensation offers to electron microscope specimen techniques. It is obvious that the usual shadowcasting technique is not quite suitable for precise studies of the two-dimensional geometry of an object, because here the density of the object is increased only on one side. The shadow-side of the object has not gained in density and there is no improvement of object definition. The rotary condensation increases the density all around the object and accordingly an improved two-dimensional definition of the object is achieved. Since the increase of density of the object or replica profile is a function of the angle between evaporating source and object, the angle has to be selected according to the surface structure of objects. In some cases, two or more evaporating sources at different angles could be used. Such a sample would be a replica of salt particles (Fig. 6), where a better particle imitation could be achieved, if one additional evaporating source at larger angle would be used. In many re-

FIG. 6. Silica replica of KCl particles, gold shadowcasting combined with rotary condensation. Negative print.

search problems, it is necessary not only to demonstrate the presence of a particle, but to reproduce the original form in such accurate detail that it can be identified. In some cases, heavy rotary condensation is necessary if greatly increased contrast is required. Such is the case, for example, when large aggregates of small elements are studied. Usual shadowcasting for such a specimen is in

some cases not suitable, because of the interfering effect of shadows with the closely packed objects themselves.

Rotary condensation does not produce any appreciable shadow and accordingly does not provide three-dimensional information. Shadowcasting, combined with rotary condensation in many instances in our experience, gives better results than either of them singly, but proper balance between the intensities of the two has to be established.

CONCLUSIONS

1. A modified silica replica technique for electron microscopy has been developed involving direct condensation of silica upon the object. Satisfactory replicas of biological membranes and various other particles have been obtained.

2. A method of rotary condensation of metallic vapors on the specimen has been developed. Compared to usual shadowcasting, a better two-dimensional definition of objects is obtained. Also, by combined use of rotary condensation and shadowcasting, a better general definition of objects is achieved than would be obtainable by either method alone.

3. Accurate reproducibility of the morphology of crystals and particles by the modified silica replica is demonstrated with electron micrographs.

The Enumeration of Fourier Phase Permutations Using Electronic Analog Machines

A. D. Booth
Birkbeck College Research Laboratory, London, England
(Received October 27, 1948)

An automatic method is proposed for recording systematically the density distributions resulting from all permutations of sign of a given set of Fourier coefficients.

It is suggested that the device could be used in conjunction with analog computers of the type now being built by Pepinski, and relay circuits for this purpose are given.

It is the purpose of this note to explain how, with simple addition, electronic two-dimensional analog computers, of the type suggested by Booth and Rivlin,[1] Rogers,[2] and Pepinski[3] can be made to record systematically the results of all permutations of sign of the given Fourier coefficients.

In general, machines of the type under consideration form the sums of the series

$$\rho(x, y) = \frac{1}{A}\sum_{h,k}(\pm)F(h, k)\cos 2\pi\left(\frac{x}{a}h + \frac{y}{b}k\right)$$

at all points in the area $(0 \leqslant x \leqslant a,\ 0 \leqslant y \leqslant b)$, and present the results either in the form of a density map or as a set of contours[3] on the face of a cathode-ray screen.

The quantities $F(h, k)$ are to be set up on a bank of potentiometers and the signs $(\pm)$ on switches; the problem is the systematic generation of $\rho(x, y)$ for all possible combinations of signs of the $F(h, k)$'s.

Of course, it is not to be supposed that any very large number of terms could be included in this enumeration and in fact, with the speeds probably available with the Pepinski machine, something like 12 planes would form an upper limit, since the 4096 permutations would take between one and two days of machine time. However, even this small number may sometimes be useful, especially if the planes chosen for examination are of large magnitude.

A possible procedure would be the following:

(1) Select the 12 largest planes and examine all permutations.

(2) Pick out the most promising groups from these.

(3) Add 12 more planes to each of the groups so chosen and examine the result of permuting these new planes.

In order to effect a complete enumeration of all possible signs of the Fourier coefficients, it is sufficient to observe that if each positive be associated with the "on" position of a switch and be called unit and similarly each negative be "off" or zero, then all combinations of sign are included in the group of binary numbers less than 2^n where n is the number of coefficients. For example, with three planes:

Signs	Binary number	Decimal equivalent
− − −	0 0 0	0
− − +	0 0 1	1
− + −	0 1 0	2
− + +	0 1 1	3
+ − −	1 0 0	4
+ − +	1 0 1	5
+ + −	1 1 0	6
+ + +	1 1 1	7

where $7 < 8 (= 2^3)$.

[1] A. D. Booth, *Fourier Technique in X-Ray Organic Structure Analysis* (Cambridge University Press, London, 1948).
[2] D. Rogers, private communication per Professor Bernal (1945).
[3] R. Pepinski, Nature 162, 22 (1948).

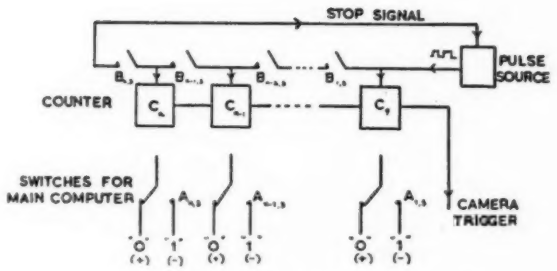


FIG. 1. Schematic of relay counter circuit.

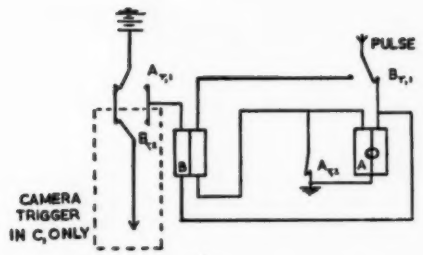


FIG. 2. Detail of counter unit C_r.

The method of causing the machine to examine these groupings is now obvious. An n stage, binary counter is set up and in each stage an on-off switch element is included. This counter is pulsed, at a rate suitable for the operation of the main computor, and the resulting screen patterns photographed. As the time for the machine to attain equilibrium may be quite long (0.1 sec. or more) a relay counter should be adequate and a suggested circuit is given in Figs. 1 and 2.

C_1-C_n (Fig. 1) represent the relay counting elements detailed in Fig. 2. A pulse source delivers voltage pulses of suitable repetition rate to these counters. In the first instance all the "carry" contacts, $B_{r,3}$, are open so that only stage C_1 reacts, the A relay closes and all $A_{1,t}$ contacts change over. $A_{1,3}$ changes the sign pattern of the main computer from 000 ---0 to 000 ----1; when things have settled down the pulse is removed from the counter and the $B_{r,t}$ contacts change. The $A_{r,t}$ contacts are unaffected so that C_1 emits a pulse to the cine-camera which photographs the screen pattern.

The next pulse finds $B_{1,3}$ closed so that both C_1 and C_2 react, C_1 returning to state zero and C_2 going to state 1.

In succeeding pulses all stages up to C_n react, and when finally C_n------ C_1 each contain unity and all $B_{r,3}$ are closed, the next pulse inhibits the pulse source and signals the completion of the operation.

Details of a suitable counter unit are given in Fig. 2. Relays A and B each have three contacts, $A_{r,1}$, $A_{r,3}$ change over $A_{r,2}$ break, $B_{r,1}$ and $B_{r,2}$ change over and $B_{r,3}$ make. The relays have two equal coils each, and these are used in opposition for A and in conjunction in B.

With the circuit in zero state, as drawn, a positive voltage pulse applied to the input causes A to react, via $B_{r,1}$. $A_{r,1}$ changes over but B does not react since both ends of its coil are at battery potential. When the pulse is removed, however, current flows through $A_{r,1}$, B, and A; this keeps A on and at the same time changes B over.

When the next pulse arrives it is steered, via $B_{r,1}$, to the other coil of B and thence to A. Since the coils of A are in opposition A goes off, however to ensure that ringing does not take place $A_{r,2}$—an early make—shorts out the coil of A and thus insures its remaining in the off condition. B still remains on, but when the pulse is removed B also returns to the off position.

It is easily seen that the camera trigger circuit (in C_1 only) operates only when the trigger pulse is removed thus giving the main computer circuit time to settle down.

Of course, if high speeds of operation of the primary machine prove possible, an electronic counter could be substituted using one of the familiar circuits,[4] however, the relay counter described above works very satisfactorily at speeds up to 50 c/sec. and will most probably prove adequate for any machine at present envisaged.

It is interesting to note that if the permuted planes are arranged in order of spacing, the cinefilm should present a more or less continuous variation of density pattern and the results of a day's machine operation could be speeded up for visualization in the reverse manner to a slow motion film.

[4] A. D. Booth, J. Inst. Electronics, p. 5 (September, 1948).

On Surface-Fitting in Three Variables

MORRIS MORDUCHOW
Polytechnic Institute of Brooklyn, Brooklyn, New York
(Received November 1, 1948)

A simple systematic method, based on a theorem stated and proven here, is presented of obtaining a formula of the form $z=f(x, y)$ from an experimental curve (Fig. 1) in which y is kept constant and a second experimental curve (Fig. 2) in which x is kept constant. Questions of the sufficiency of the original data are automatically answered by this method. Several examples are given to illustrate the procedure explicitly.

INTRODUCTION

LET it be required to obtain an empirical formula, within a given range of values of x and y, of the form

$$z=f(x, y) \tag{1}$$

from a set of data given by two single curves of the form of Fig. 1 and Fig. 2. For each figure, one of the variables (e.g., y in Fig. 1) is kept constant, while the other is permitted to vary. The aim of this paper is to present a simple systematic method, sufficiently general for many practical purposes, of obtaining a relation of the form (1) from two curves such as Figs. 1 and 2. This method automatically reveals cases in which Figs. 1 and 2 may be insufficient to determine uniquely a relation of the form (1), and it determines, moreover, the minimum additional data required in such cases to obtain a unique formula.

It will be assumed that, within the desired range, Figs. 1 and 2 are typical, so that they will have the same form regardless of the values $y=y_0$ or $x=x_0$ at which either variable is kept fixed.

BASIC THEOREM

The following theorem is the basis of the method presented here.

Theorem.—If (a) the curve in either Fig. 1 or Fig. 2 implies a formula of the form $z=f_1(x)\cdot f_2(y)$, *or* if (b) the curves in both Figs. 1 and 2 are polynomials, then the *most general* form of the function $z=f(x, y)$ satisfying the data in these figures can be obtained by multiplying $\phi(x, a_1, a_2, \cdots, a_p)$ by $\psi(y, b_1, b_2, \cdots, b_q)$, where $z=\phi$ is a formula fitting Fig. 1, while $z=\psi$ is a formula fitting Fig. 2.

Proof: (a) In Fig. 1 let $z=a_1\phi(x)$, and in Fig. 2 let $z=b_1\psi(y)$, where ϕ and ψ are definite functions of x and of y, respectively. Then, in general, the coefficient a_1 will be a function of y, while the coefficient b_1 will be a function of x. These functions must be obtained from the condition that Figs. 1 and 2 imply exactly the same formula for z, i.e., that

$$a_1(y)\cdot\phi(x)\equiv b_1(x)\cdot\psi(y). \tag{2}$$

Equation (2) is an identity-equation, which can be solved for the unknown functions $a_1(y)$ and $b_1(x)$ by separation of variables; thus

$$(a_1(y)/\psi(y))\equiv(b_1(x)/\phi(x))=k,$$

where k is an arbitrary constant. Thus,

$$a_1(y)=k\psi(y), \quad b_1(x)=k\phi(x).$$

Hence, the required formula must be of the form

$$z=k\phi(x)\cdot\psi(y), \quad \text{Q.E.D.}$$

(b) Let the curve in Fig. 1 be of the form

$$z=\sum_{i=0}^{p} a_i x^i, \tag{3}$$

while the curve in Fig. 2 is of the form

$$z=\sum_{j=0}^{q} b_j y^j. \tag{4}$$

Then, following the same reasoning as in (a), the coefficients a_i and b_j may, in general, be considered as functions of y and of x, respectively. These functions, moreover, must be such that they satisfy the identity-equation

$$\sum_{i=0}^{p} a_i(y)x^i\equiv\sum_{j=0}^{q} b_j(x)y^j. \tag{5}$$

Differentiating (5) with respect to $y(q+1)$ times, one obtains the identity-equation

$$\sum_{i=0}^{p} a_i^{[q+1]}(y)x^i\equiv 0. \tag{6}$$

From (6) it follows that $a_i^{[q+1]}(y)=0$, and that $a_i(y)$ must therefore be of the form

$$a_i(y)=P_{qi}(y), \tag{7}$$

where $P_{qi}(y)$ is a polynomial in y of degree q.

Similarly, by differentiating (5) with respect to $x(p+1)$ times, it can be shown that $b_j(x)$ must have the form

$$b_j(x)=P_{pj}(x), \tag{8}$$

where $P_{pj}(x)$ is a polynomial in x of degree p.

The most general form of the function $z=f(x, y)$ represented by Figs. 1 and 2 here can now be obtained by putting (7) into (3), or (8) into (4). From the results thus obtained, it is seen that one arrives at a form which is exactly the same as that which would be obtained by multiplying (3) by (4) directly, Q.E.D.

Part (b) of the above theorem may have special theoretical, as well as practical, interest, since in a given region any continuous curve can, in principle, be represented to any degree of approximation by a polynomial. In principle, therefore, this theorem would cover practically all possible cases.

The method of proof of the above theorem, namely, that of deriving and solving an identity-equation, may be considered as a general method of surface-fitting in three variables, in cases which do not satisfy the conditions of the theorem. It is believed, however, that most practical cases will satisfy these conditions, and it is, in fact, suggested that in the actual procedure of curve-fitting, special attempts be always first made to obtain formulas of the type (a) or (b) satisfying Figs. 1 and 2.

After the form of the formula $z=f(x, y)$ has been determined, the values of the coefficients there can be obtained from the numerical values of the original coefficients a_i and b_j at $y=y_0$ and at $x=x_0$, respectively. This usually leads to a simple set of linear algebraic equations, which, in addition to furnishing a complete solution, answers questions of consistency and of sufficiency of the original data.

Certain cases other than those covered directly by (a) and (b) can be readily transformed into case (a) or case (b) (cf. examples 3 and 4 below).

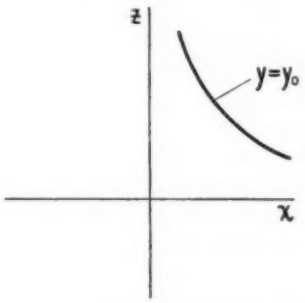


FIG. 1.

EXAMPLES*

1. Derivation of General Gas Law from Boyle's Law and Charles' Law

Let Fig. 1 be a $p-V$ diagram, indicating

$$pV=a_1 \quad \text{for} \quad T=T_0, \tag{9}$$

while Fig. 2 is a $p-T$ diagram, indicating

$$(p/T)=b_1 \quad \text{for} \quad V=V_0. \tag{10}$$

* The subscript "0" will be used here to denote the value of a_i when $y=y_0$ and of b_j when $x=x_0$, e.g., $a_{20}\equiv(a_2)_{y=y_0}$.

Writing (9) in the form $p=(a_1/V)$, and (10) in the form $p=b_1T$, it is seen that (9) and (10) satisfy condition (a) of the theorem. Hence, the general relation between p, V, and T must be of the form

$$p=k(T/V), \quad \text{or} \quad pV=kT. \tag{11}$$

2. Two Straight Lines

Let Fig. 1 be of the form

$$z=a_1+a_2x, \tag{12}$$

while Fig. 2 is of the form

$$z=b_1+b_2y. \tag{13}$$

Since (12) and (13) are polynomials, they satisfy condition (b) of the theorem. Hence, the most general form for $z=f(x, y)$ is obtained by multiplying (12) by (13). The form thus obtained is

$$z=k_1+k_2x+k_3y+k_4xy, \tag{14}$$

where the "k" are constants. The numerical values of k are determined from the following relations:

$$\left.\begin{aligned} a_{10}+a_{20}x &\equiv k_1+k_2x+k_3y_0+k_4xy_0, \\ \therefore \quad k_1+k_3y_0&=a_{10} \\ k_2+k_4y_0&=a_{20} \end{aligned}\right\}. \tag{15}$$

Moreover,

$$b_{10}+b_{20}y\equiv k_1+k_2x_0+k_3x+k_4x_0y;$$

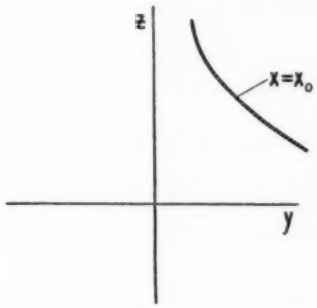


FIG. 2.

therefore,

$$\left.\begin{aligned} k_1+k_2x_0&=b_{10} \\ k_3+k_4x_0&=b_{20} \end{aligned}\right\}. \tag{16}$$

(15) and (16) are 4 linear equations in the 4 unknowns, k_1, k_2, k_3, and k_4. It will be found, however, that these equations are not all independent, so that one can at best determine three of the unknowns in terms of the fourth (say k_4). Subsequent substitution into (14) then gives:

$$z=(b_{10}-a_{20}x_0)+a_{20}x+b_{20}y+k_4(x-x_0)(y-y_0). \tag{17}$$

Equation (17) is the desired formula. The data is here evidently insufficient to determine a unique formula for z, since k_4 is arbitrary. The value of k_4 can, however, be determined by performing only a

single additional experiment to obtain a set of values of (x, y, z) not on either of the two original straight lines.**

As in all cases, the condition that the original data be consistent can be obtained here simply by putting $x=x_0$ into Eq. (12) and $y=y_0$ into Eq. (13). The following relation is thus obtained:

$$a_{10}+a_2x_0=b_{10}+b_{20}y_0. \tag{18}$$

If the original data does not satisfy Eq. (18), then the data is inconsistent.***

3. Two Exponentials

Suppose Figs. 1 and 2 can be represented, respectively, by formulas of the forms

$$z=a_1e^{a_2x}, \quad z=b_1e^{b_2y}.$$

These do not directly satisfy either of the conditions of the basic theorem presented here. If, however, one writes

$$\log z=\log a_1+a_2x, \quad \log z=\log b_1+b_2y,$$

then one obtains forms satisfying condition (b) of the theorem. Hence the most general form of $\log z$ will be

$$\log z=k_1+k_2x+k_3y+k_4xy. \tag{19}$$

Thus, the most general form for z in this case will be

$$z=\exp(k_1+k_2x+k_3y+k_4xy), \tag{20}$$

where the k are constants which can be determined in a manner similar to that shown in example 2.

4. Parabola and Exponential

Suppose Fig. 1 can be fitted by the formula

$$z=3x^2 \quad \text{for} \quad y=1, \tag{21}$$

while Fig. 2 can be fitted by the formula

$$z=4\exp(1.5y) \quad \text{for} \quad x=2.44. \tag{22}$$

These forms may be considered as special cases of the more general forms

$$z=a_1x^{a_2} \quad \text{and} \quad z=b_1\exp(b_2y),$$

respectively, where the "a" are functions of y and the "b" are functions of x. As in example 3, these forms do not satisfy the conditions of the basic theorem. However, by writing these forms as

$$\log z=\log a_1+a_2\log x \quad \text{and} \quad \log z=\log b_1+b_2y,$$

it is seen that condition (b) of the theorem is satisfied for $\log z$ as a function of $\log x$ and of y, so that the most general form for $\log z$ must, according to the theorem, be

$$\log z=k_1+k_2y+k_3\log x+k_4y\log x, \tag{23}$$

where the k are constants. The numerical values of the k can be obtained from the original data by the procedure shown in example 2. It may be noted from (21) and (22) that the following numerical values are given here:

$$a_{10}=3, \quad a_{20}=2, \quad b_{10}=4, \quad b_{20}=1.5, \quad x_0=2.44, \quad y_0=1.$$

The final result for z is found to be

$$z=3x^2\exp\{(y-1)[1.5(\log x/\log 2.44)+k_2(1-(\log x/\log 2.44))]\}. \tag{24}$$

In (24), k_2 is an arbitrary constant. The data is here evidently insufficient to determine a unique formula. However, only one additional set of data (x, y, z) not on either of the two original curves is required to obtain the value of k_2, and thus to obtain a unique formula.

CONCLUSIONS

In the actual process of fitting a formula of the type (1) to a set of data given in the form of Figs. 1 and 2, attempts should always be made first to fit curves to both Figs. 1 and 2 satisfying the conditions [either (a) or (b)] of the basic theorem presented here. It is believed that in many actual cases such attempts will be successful. A simple transformation of variables may sometimes be necessary for this purpose. Application of the theorem given here then permits a rapid determination of the desired formula (1), either in complete form, or in as complete form as possible with the data given in Figs. 1 and 2. If formulas for Figs. 1 and 2 which satisfy the conditions of the theorem cannot be obtained, then an identity-equation may have to be solved.

In view of the theorem proven here, figures of the form of 1 and 2 are especially convenient for obtaining an empirical relation of the form $z=f(x, y)$, since the problem of fitting a surface in two variables is here reduced to fitting two single curves in one variable each.

The author wishes hereby to express his gratitude for the very kind and invaluable suggestions and aid given him by Professor Ronald M. Foster in connection with this paper.

** For checking purposes, more than one such set should, of course, be obtained.

*** Equation (18) could also be automatically derived by solving Eqs. (15) and (16) for the k, since (18) is the condition that these four equations be consistent.

Reflection and Transmission of Electromagnetic Waves by a Spherical Shell*

HERBERT B. KELLER AND JOSEPH B. KELLER
Washington Square College Mathematics Research Group, New York University, New York, New York
(Received November 5, 1948)

Exact, explicit expressions are obtained for the field of a radiating periodic point dipole located in medium 1 at the center of a spherical shell of medium 2, which is bounded by medium 3. The values of ϵ, μ, σ and the shell radii are arbitrary. The reflected and transmitted fields are examined in various special cases. When the shell radii become infinite, the fields agree with those obtained by R. K. Luneberg for the case of plane waves normally incident on a flat plate of medium 2 between media 1 and 3. When the radii are infinite and media 2 and 3 identical, the Fresnel formulas for normal incidence of plane waves on a plane interface between two half-infinite media are obtained. When the radii are finite, media 2 and 3 identical and $\mu_1=\mu_3$, the results of C. T. Tai are obtained. For finite radii, media 1 and 3 identical, and shell thickness small compared to a wave-length, the fields check the approximate results of J. B. Keller for the fields reflected and transmitted by a thin shell of any shape, when his results are specialized to the present case.

INTRODUCTION

THE electromagnetic field due to a radiating periodic point dipole in a homogeneous isotropic medium of infinite extent is well known.[1] In this paper we intend to find the modification of the field when the source is at the center of a medium bounded by a spherical shell of different material, with a third medium of infinite extent external to the shell. It is expected that part of the field will be reflected from the shell and part transmitted. These reflected and transmitted fields are of particular interest. This problem is solved exactly and the solution is examined in special cases, in some of which it reduces to known results. In addition, the solution is used to test the accuracy of a recent approximate calculation of the reflection and transmission by an arbitrarily curved thin shell.[2]

FORMULATION OF THE PROBLEM

A point dipole with axis in the z-direction is at the origin of coordinates in medium 1, with constants μ_1, ϵ_1, and σ_1, and bounded at $r=r_a$ by medium 2. This medium extends from $r=r_a$ to $r=r_b$ and has the constants μ_2, ϵ_2, and σ_2. Bounding the shell at $r=r_b$ is medium 3 of infinite extent and with constants μ_3, ϵ_3, and σ_3.

The electromagnetic field vectors E and H must satisfy Maxwell's equations[3] in each medium and their tangential components must be continuous at the two interfaces between the shell and the surrounding media.[4] In addition the field must vanish at infinity and also contain the incident field. The time dependence of the fields is assumed to be of the form $e^{i\omega t}$, where ω is the angular frequency of the dipole, no volume charge is present, and μ, ϵ, and σ are constant in each region.

USE OF VECTOR POTENTIAL

We introduce the vector potential A which satisfies

$$\nabla^2 A+(\omega^2\mu\epsilon'/c^2)A=0. \tag{1}$$

It is easily verified that the field vectors will satisfy Maxwell's equations if we define them in terms of the vector potential, using Gaussian units, as[5]

$$\begin{aligned} H&=\nabla\times A \\ E&=(c/i\omega\epsilon')\nabla(\nabla\cdot A)-(i\omega\mu/c)A \end{aligned} \tag{2}$$

where

$$\epsilon'=\epsilon-(4\pi i\sigma/\omega). \tag{3}$$

For a point dipole with axis in the z-direction, in a medium of infinite extent, the solution of Eq. (1) is found to be a vector with a single component[6]

$$A_z=A_0(e^{-ikr}/r) \tag{4}$$

where A_0 is a constant proportional to the dipole strength, and

$$k=\omega/c(\mu\epsilon')^{\frac{1}{2}}. \tag{5}$$

SOLUTION OF THE PROBLEM

In the present case we assume the solution of Eq. (1) to be a vector with the single non-zero component:

$$\begin{aligned} A_z&=A_0\frac{e^{-ik_1r}}{r}+A_1\frac{e^{ik_1r}}{r} && 0<r<r_a \\ A_z&=A_2\frac{e^{-ik_2r}}{r}+A_3\frac{e^{ik_2r}}{r} && r_a<r<r_b \\ A_z&=A_4\frac{e^{-ik_3r}}{r} && r_b<r. \end{aligned} \tag{6}$$

* This work was performed under contract with the Army Air Forces, Watson Laboratories, Red Bank, New Jersey.

[1] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941) pp. 436–7; S. Ramo and J. R. Whinnery, *Fields and Waves in Modern Radio* (John Wiley & Sons, Inc., New York, 1944) p. 430.

[2] Joseph B. Keller, "Reflection and transmission of electromagnetic waves by thin curved shells," Thesis, New York University, April 1948.

[3] S. Ramo, and J. R. Whinnery, see reference 1, p. 154.

[4] J. A. Stratton, see reference 1, pp. 483–4.

[5] S. Ramo, and J. R. Whinnery, see reference 1, p. 165.

[6] S. Ramo and J. R. Whinnery, see reference 1, p. 430.

In this solution the constant A_0 is presumed to be known since it is determined by the dipole strength, i.e., the outgoing wave in medium 1 is assumed to be given by Eq. (4), and the constants A_1, A_2, A_3, A_4 are to be determined. From Eqs. (6) it can be seen that the assumed solution contains the incident field (4) and vanishes at infinity. By substitution one finds that (6) is actually a solution of Eq. (1), and therefore the field components, which are obtained by using (6) in Eqs. (2), will satisfy Maxwell's equations in each region.

The four remaining conditions which the solution must satisfy are the requirements that the tangential components of E and H must be continuous at $r=r_a$ and $r=r_b$. When the solution, Eq. (6), is used in Eqs. (2) and the appropriate field components are equated at $r=r_a$ and $r=r_b$, four inhomogeneous linear algebraic equations are obtained. These equations may be solved for A_1, A_2, A_3, and A_4, which completes the solution of the problem. The vector potential of the reflected and transmitted fields are then found to be given, respectively, by the following expressions:

$$A_1\frac{e^{ik_1r}}{r}=-A_0\frac{\exp(-i[k_1r_a+k_1(r_a-r)])}{r}\left[\frac{e^{ik_2h}F\cdot M+e^{-ik_2h}J\cdot P}{e^{ik_2h}F\cdot G+e^{-ik_2h}J\cdot L}\right] \tag{7}$$

$$A_4\frac{e^{-ik_3r}}{r}=A_0\epsilon_2'\epsilon_3'\frac{\exp(-i[k_1r_a+k_3(r-r_b)])}{r}\left[\frac{Q\cdot S}{e^{ik_2h}F\cdot G+e^{-ik_2h}J\cdot L}\right] \tag{8}$$

where the new symbols are defined by

$$\begin{aligned}
F&=\left\{\epsilon_3'\left[ik_3+\frac{1}{r_b}\right]\left[ik_2^2-\frac{k_2}{r_b}-\frac{i}{r_b^2}\right]-\epsilon_2'\left[-ik_2+\frac{1}{r_b}\right]\left[ik_3^2+\frac{k_3}{r_b}-\frac{i}{r_b^2}\right]\right\}\\
G&=\left\{\epsilon_1'\left[-ik_1+\frac{1}{r_a}\right]\left[ik_2^2+\frac{k_2}{r_a}-\frac{i}{r_a^2}\right]-\epsilon_2'\left[ik_2+\frac{1}{r_a}\right]\left[ik_1^2-\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]\right\}\\
J&=\left\{\epsilon_3'\left[ik_3+\frac{1}{r_b}\right]\left[ik_2^2+\frac{k_2}{r_b}-\frac{i}{r_b^2}\right]-\epsilon_2'\left[ik_2+\frac{1}{r_b}\right]\left[ik_3^2+\frac{k_3}{r_b}-\frac{i}{r_b^2}\right]\right\}\\
L&=\left\{\epsilon_2'\left[-ik_2+\frac{1}{r_a}\right]\left[ik_1^2-\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]-\epsilon_1'\left[-ik_1+\frac{1}{r_a}\right]\left[ik_2^2-\frac{k_2}{r_a}-\frac{i}{r_a^2}\right]\right\}\\
M&=\left\{\epsilon_1'\left[ik_1+\frac{1}{r_a}\right]\left[ik_2^2+\frac{k_2}{r_a}-\frac{i}{r_a^2}\right]-\epsilon_2'\left[ik_2+\frac{1}{r_a}\right]\left[ik_1^2+\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]\right\}\\
P&=\left\{\epsilon_2'\left[-ik_2+\frac{1}{r_a}\right]\left[ik_1^2+\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]-\epsilon_1'\left[ik_1+\frac{1}{r_a}\right]\left[ik_2^2-\frac{k_2}{r_a}-\frac{i}{r_a^2}\right]\right\}\\
Q&=\left\{\left[-ik_2+\frac{1}{r_b}\right]\left[ik_2^2+\frac{k_2}{r_b}-\frac{i}{r_b^2}\right]-\left[ik_2+\frac{1}{r_b}\right]\left[ik_2^2-\frac{k_2}{r_b}-\frac{i}{r_b^2}\right]\right\}\\
S&=\left\{\left[ik_1+\frac{1}{r_a}\right]\left[ik_1^2-\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]-\left[-ik_1+\frac{1}{r_a}\right]\left[ik_1^2+\frac{k_1}{r_a}-\frac{i}{r_a^2}\right]\right\}\\
h&=r_b-r_a.
\end{aligned} \tag{9}$$

Corresponding expressions are obtained for the vector potential in medium 2.

DISCUSSION OF THE SOLUTION

The vector potential of the reflected and transmitted fields given by Eqs. (7) and (8), assume simpler forms in special cases, some of which will now be examined.

I. When $h=0$ or $r_a=r_b$ the shell of medium 2 is absent. In this case Eqs. (7) and (8) become

$$A_1\frac{e^{ik_1r}}{r}=A_0\frac{\exp(-i[k_1r_a+k_1(r_a-r)])}{r}$$

$$\times\left\{\frac{\left[k_1k_3(\epsilon_3'k_1-\epsilon_1'k_3)-\frac{(k_1+k_3)}{r_a^2}(\epsilon_3'-\epsilon_1')\right]-i\left[\frac{(k_1+k_3)}{r_a}(\epsilon_3'k_1-\epsilon_1'k_3)-\frac{(\epsilon_3'-\epsilon_1')}{r_a^3}\right]}{\left[-k_1k_3(\epsilon_3'k_1+\epsilon_1'k_3)+\frac{(k_3-k_1)}{r_a^2}(\epsilon_3'-\epsilon_1')\right]-i\left[\frac{(k_3-k_1)}{r_a}(\epsilon_3'k_1+\epsilon_1'k_3)-\frac{(\epsilon_3'-\epsilon_1')}{r_a^3}\right]}\right\} \tag{10}$$

$$A_4\frac{e^{-ik_3r}}{r}=A_0\frac{\exp(-i[k_1r_a+k_3(r-r_a)])}{r}$$

$$\times\left\{\frac{-2\epsilon_3'k_1^3}{\left[-k_1k_3(\epsilon_3'k_1+\epsilon_1'k_3)+\frac{(k_3-k_1)}{r_a^2}(\epsilon_3'-\epsilon_1')\right]-i\left[\frac{(k_3-k_1)}{r_a}(\epsilon_3'k_1+\epsilon_1'k_3)-\frac{(\epsilon_3'-\epsilon_1')}{r_a^3}\right]}\right\}. \tag{11}$$

If, in addition, r_a becomes infinite while $r-r_a$ remains finite and $A_0(e^{-ik_1r_a}/r)$ approaches a limit A_0', the problem reduces to that of a plane wave normally incident on a plane interface between two half-infinite media. In this case, writing $r-r_a=x$, Eqs. (10) and (11) reduce to

$$A_1\frac{e^{+ik_1r}}{r}=A_0'e^{ik_1x}\left\{\frac{\epsilon_1'k_3-\epsilon_3'k_1}{\epsilon_1'k_3+\epsilon_3'k_1}\right\} \tag{12}$$

$$A_4\frac{e^{-ik_3r}}{r}=A_0'e^{-ik_3x}\left\{\frac{2\epsilon_3'\frac{k_1^2}{k_3}}{\epsilon_1'k_3+\epsilon_3'k_1}\right\}. \tag{13}$$

Equations (12) and (13) are the well-known Fresnel formulas for normal incidence[7] and Eqs. (10) and (11) are generalizations of these formulas. If $\mu_1=\mu_3$ Eqs. (10) and (11) reduce to those obtained by C. T. Tai[8] for this special case.

II. When r_b and r_a become infinite while $r_b-r_a=h$ and $r-r_a=x$ remain finite and $A_0(e^{-ik_1r_a}/r)$ approaches a limit A_0', the problem reduces to that of plane waves normally incident on a flat plate separating two different half-infinite media. In this case Eqs. (7) and (8) become:

$$A_1\frac{e^{ik_1r}}{r}=A_0'e^{ik_1x}\cdot\left\{\frac{\epsilon_2'k_2(\epsilon_1'k_3-\epsilon_3'k_1)\cos k_2h+i(\epsilon_1'\epsilon_3'k_2^2-\epsilon_2'^2k_1k_3)\sin k_2h}{\epsilon_2'k_2(\epsilon_1'k_3+\epsilon_3'k_1)\cos k_2h+i(\epsilon_1'\epsilon_3'k_2^2+\epsilon_2'^2k_1k_3)\sin k_2h}\right\}, \tag{14}$$

$$A_4\frac{e^{-ik_3r}}{r}=A_0'e^{-ik_3(x-h)}\cdot\left\{\frac{2\epsilon_2'\epsilon_3'k_1^2\frac{k_2}{k_3}}{\epsilon_2'k_2(\epsilon_1'k_3+\epsilon_3'k_1)\cos k_2h+i(\epsilon_1'\epsilon_3'k_2^2+\epsilon_2'^2k_1k_3)\sin k_2h}\right\}. \tag{15}$$

Equations (14) and (15) are the same as those obtained by R. K. Luneberg for this case.[9]

III. If $h/r_a\ll 1$ and $|k_2h|\ll 1$ then the Hertz vectors given by Eqs. (7) and (8) are accurately represented by the first few terms in their Taylor's expansions in powers of h. The first terms of these expansions are

[7] J. A. Stratton, see reference 1, p. 492.
[8] C. T. Tai, "Radiation of a Hertzian dipole immersed in a dissipative medium," Harvard University, Cruft Laboratory, Research Report No. 21.
[9] R. K. Luneberg, "The propagation of electromagnetic plane waves in plane parallel layers," Washington Square College Mathematics Research Group, Research Report No. 172–3, June 1947, pp. 48–9.

given below for the case in which medium 3 is the same as medium 1, and $|1/k_1r_a|\ll 1$.

$$A_1\frac{e^{ik_1r}}{r}=A_0\frac{\exp(-i[2k_1r_a-k_1r])}{r}\frac{ik_2h}{2}\left\{\frac{\epsilon_1'k_2}{\epsilon_2'k_1}-\frac{\epsilon_2'k_1}{\epsilon_1'k_2}\right\} \tag{15}$$

$$A_4\frac{e^{-ik_1r}}{r}=A_0\frac{e^{-ik_1r}}{r}\left\{1+ik_1h-\frac{ik_2h}{2}\left[\frac{\epsilon_1'k_2}{\epsilon_2'k_1}+\frac{\epsilon_2'k_1}{\epsilon_1'k_2}\right]\right\}. \tag{16}$$

By using (15) and (16) in Eqs. (2) we obtain the reflected and transmitted field components expanded in powers of h. These are, when $1/|k_1r|\ll 1$,

$$H_{\phi R}=A_1\frac{e^{ik_1r}}{r}(-ik_1\sin\theta) \tag{17}$$

$$H_{\phi T}=A_4\frac{e^{-ik_1r}}{r}(ik_1\sin\theta) \tag{18}$$

$$E_{\theta R}=A_1\frac{e^{ik_1r}}{r}\left(\frac{ick_1^2}{\omega\epsilon_1'}\sin\theta\right) \tag{19}$$

$$E_{\theta T}=A_4\frac{e^{-ik_1r}}{r}\left(\frac{ick_1^2}{\omega\epsilon_1'}\sin\theta\right) \tag{20}$$

$$E_{rR}=A_1\frac{e^{ik_1r}}{r}\left(-\frac{2ck_1}{\omega\epsilon_1'r}\cos\theta\right) \tag{21}$$

$$E_{rT}=A_4\frac{e^{-ik_1r}}{r}\left(\frac{2ck_1}{\omega\epsilon_1'r}\cos\theta\right) \tag{22}$$

$$H_r=H_\theta=E_\phi=0. \tag{23}$$

In these equations the letters R and T in the subscripts denote reflected and transmitted field components, respectively.

COMPARISON WITH AN APPROXIMATE SOLUTION

An approximate solution of the reflection and transmission of plane or spherical electromagnetic waves from an arbitrarily curved thin shell has recently been given by J. B. Keller.[2] The results, when specialized to the case of plane waves incident on a thin flat plate, agree with those of R. K. Luneberg.[9] In order to obtain another check on the approximate solution we will specialize it to the case of a point source at the center of a spherical shell. The fields obtained in this way should agree with those given by Eqs. (17)–(23) above.

When Eq. (45) in the paper by J. B. Keller and the corresponding equation for H are specialized to this case it is found that the reflected and transmitted H_r, H_θ, and E_ϕ are all zero. We also find that the reflected and transmitted H_ϕ and E_θ are given by Eqs. (17)–(20). The reflected and transmitted E_r components become:

$$E_{rR}=A_1\frac{e^{ik_1r}}{r}\left(-\frac{2ck_1}{\omega\epsilon_1'r}\cos\theta\right)\frac{r}{r_a} \tag{24}$$

$$E_{rT}=\left[A_0+(A_4-A_0)\frac{r}{r_a}\right]\frac{e^{-ik_1r}}{r}\left(\frac{2ck_1}{\omega\epsilon_1'r}\cos\theta\right). \tag{25}$$

We thus see that the approximate solution agrees with the special form of the exact solution given by Eqs. (17)–(23) in all components except E_{rR} and E_{rT}. Since Eqs. (17)–(22) are valid only when $|k_1r_a|\gg 1$ and $|k_1r|\gg 1$, the approximate solution can give good results only under the same conditions. The first condition was to be expected from the derivation of the approximate solution, which assumed that the source was many wave-lengths from the obstacle. In that determination it was also assumed that there was no singularity in the reflected field. In the special case here considered there is a singularity at $r=0$ and hence it is to be expected that the result will hold only at a distance of several wave-lengths from the singularity, i.e., $|k_1r|\gg 1$.

In the case of the components E_{rR} and E_{rT}, Eqs. (24) and (25) become identical with Eqs. (21) and (22) when $r=r_a$, but differ for other values of r. The error in the radial dependence of E_r is due to the sensitive dependence of E_r on the method of evaluating an integral occurring in the approximate solution.

From the above comparison of the approximate and exact solutions it can be seen that the approximate solution is correct except for the longitudinal field component E_r, and except near the singularity of the reflected field. Since the occurrence of singularities in reflected fields is uncommon, as is also the occurrence of longitudinal components, the comparison indicates that the approximate solution should usually be acceptable.

Microwave Antennas and Dielectric Surfaces*

R. M. REDHEFFER
Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, Massachusetts
(Received November 10, 1948)

For a plane dielectric sheet near a receiving antenna the power varies sinusoidally with distance, having a half-wave period and an average proportional to the transmission coefficient. This is explained by assuming an equivalent antenna reflection, an assumption checked by special experiments. The theory includes phase and arbitrary incidence, for which the wave may be elliptically polarized. The reciprocal of received amplitude *vs.* polarization then gives a similar ellipse with 90° shift in orientation. Corresponding theory for transmitters shows that the reflection varies sinusoidally with half-wave period and with average proportional to the reflection coefficient. For a distant sheet there is inverse-distance attenuation. The effect of initial antenna mismatch is fully investigated. For any cylindrical surface, including corrugations or a plane, amplitude reflection varies with angle as the secondary power pattern. With a paraboloidal antenna of radius a at wave-length λ, the reflection from a circular cylinder of radius ρ and reflection coefficient R is $6R(\lambda\rho)^{\frac{1}{2}}/13a$; that from a narrow strip of width $2aq$ is $(7Rq/80)(15-10q^2+3q^4)$; that from a narrow strip distant aq' from the axis is proportional to $R(1-q'^2)^4$. If the initial mismatch gives a reflection r, and if the surfaces make an angle θ with the axis, the above are multiplied by $(1-r^2)\exp(4\theta^2\ln 2/w)$ where w is the antenna pattern half-width. Corresponding results are given for corrugated surfaces and series of strips. Applications to radomes are discussed, and to pressurizing seal design and tolerances. All results are extensively verified experimentally.

I. INTRODUCTION

THIS article describes the effect of a dielectric sheet in front of a transmitting or receiving microwave antenna. Such questions have importance for determining transmission and reflection coefficients, for certain methods of measuring dielectric constants, for pressurizing, and for the theory of antenna housings or *radomes*. Also the interference effects, which are somewhat different from those usually encountered for Fresnel interference in optics, are believed of interest in their own right. We present an elementary theory based on simplifying assumptions; these assumptions are verified by experiments explicitly directed to that end; and the main diffraction effects are interpreted in terms of the theory.

II. RECEIVING ANTENNA AND PLANE SHEET

1. Simple Theory

In the arrangement of Fig. 1a it is found experimentally that the received power depends on the distance to the sheet, as shown in Fig. 2. To explain this variation it was suggested by T. J. Keary that the antenna itself has an equivalent reflection coefficient of its own, which we shall denote by ρ, so that the arrangement of Fig. 1a is regarded as analogous to that of Fig. 1b. The quantities t, τ, r, and ρ are the amplitude transmission and reflection coefficients for the first object, corresponding to the antenna, while T and R are those for the second, corresponding to the dielectric sheet. The over-all reflection of antenna-plus-sheet, as it would be measured in the line with the antenna used as a transmitter, is designated by ω.

The validity of replacing Fig. 1a by 1b is discussed,[1] and it is found that the equivalent coefficients r, ω may be quite properly introduced, and that t, τ are probably permissible too if we do not assume $t=\tau$. The coefficients R, T are likewise well-defined, and for ordinary aperture illuminations one may use their plane-wave values with negligible error. The equivalent reflection for the antenna represents a more far-reaching assumption, however, and is separately discussed below.

To the extent that Fig. 1a may be replaced by 1b we obtain, without further approximation,

$$P=\frac{P_0T^2}{1-2\rho R\cos 4\pi d/\lambda+\rho^2R^2} \tag{1}$$

for P, the received power with the dielectric sheet, in terms of P_0, the power without it, and the distance d to the sheet from some suitable origin. This equation may be derived by two methods given

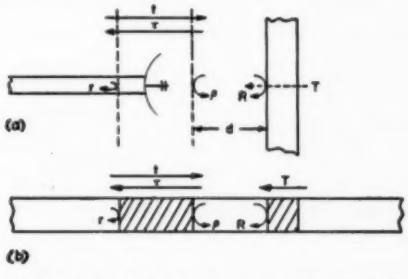


FIG. 1. Definition of equivalent coefficients. (a) Actual arrangement; (b) Simplified equivalent arrangement.

* This work has been supported in part by the Signal Corps, the Air Materiel Command, and ONR.

[1] R. M. Redheffer, "The interaction of microwave antennas with dielectric sheets," R. L. Report 483–18 (March 1, 1946).

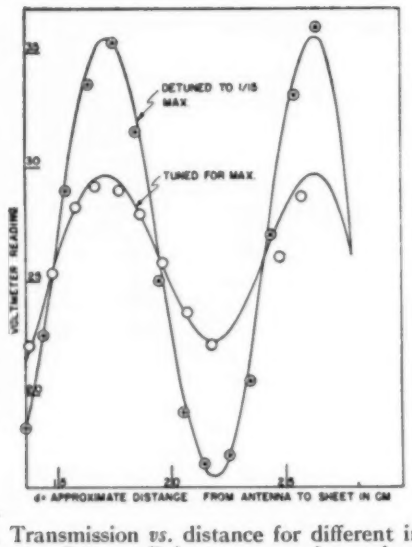


FIG. 2. Transmission *vs.* distance for different initial mismatches at $\lambda=9.1$ cm. Points are experimental, normalized for same reading without dielectric sheet; curves are best-fitting theoretical.

previously.[2] It predicts that the received power will vary periodically with period $\lambda/2$ in d, and that it will be nearly sinusoidal if ρR is sufficiently small. The first property is verified accurately, the second approximately, in all curves hitherto obtained (cf. Fig. 2).

If M^2 and m^2 are the maximum and minimum values of the received power as the position of the sheet is changed, we find

$$T=2Mm/(P_0)^{\frac{1}{2}}(M+m),$$
$$T^2\cong Mm/P_0 \quad \text{(for slide rule calculation)} \quad (2)$$

to give the transmission coefficient of the dielectric sheet. These relations are independent of ρR, and form the basis for free-space measurement of transmission. They have been extensively verified in practice; examples showing a comparison with theory are given in Fig. 3.

2. The Equivalent Reflection Coefficient of the Antenna

The validity of assuming an equivalent reflection ρ for the antenna is investigated as follows. First use Eq. (1) to obtain

$$R=(M-m)/(M+m) \quad (3)$$

with M and m as above defined. This relation was verified in reference 2, where an experimental plot of $(M-m)/(M+m)$ *vs.* R was found, within the error of measurement, to be a straight line through the origin.

Another check on the nature of ρ can be found by use of a dielectric sheet. Thus, if the antenna truly has an equivalent reflection in the sense of Fig. 1b, then the reflection ρ of sheet-plus-antenna should obey the usual law for reflection of two sheets *vs.* separation, as the distance is changed. This was also verified within the experimental error, and the value of ρ for optimum fit was substantially the same as that found by the slope of the curve in the preceding paragraph. The method of measuring ρ, must consist, incidentally, in adding a second sheet beyond the first and taking M and m as the position of this second sheet is changed. More conventional methods of measurement—e.g., use of an exploring dipole—are not relevant to the present theory.

The receiving antenna has been hitherto assumed polarized in the same direction as the incident wave. More generally we may resolve the incident wave into two components, one parallel and one perpendicular to the receiver polarization. If the perpendicular component retains its direction of polarization after successive reflections between antenna and sheet, then it may be neglected altogether; and we thus conclude that the ratio M/m should not depend on the direction of receiver polarization. This conclusion was experimentally verified.

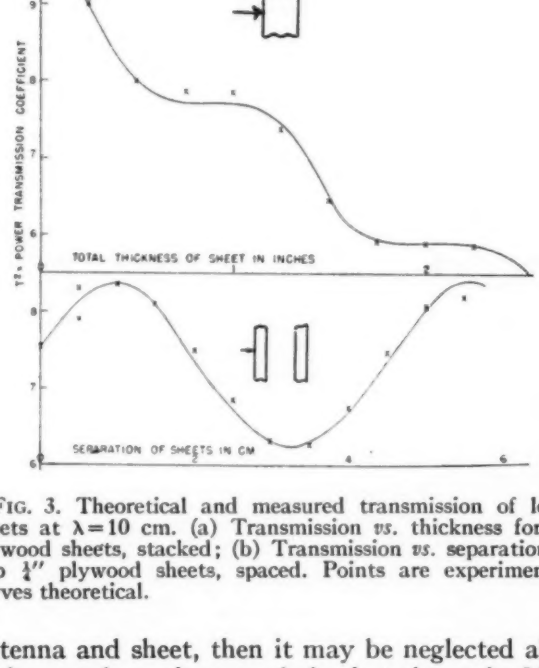


FIG. 3. Theoretical and measured transmission of lossy sheets at $\lambda=10$ cm. (a) Transmission *vs.* thickness for $\frac{1}{4}''$ plywood sheets, stacked; (b) Transmission *vs.* separation of two $\frac{1}{4}''$ plywood sheets, spaced. Points are experimental, curves theoretical.

Finally, one may add a tuner behind the antenna to permit variation of the reflection coefficient r in Fig. 1. It is probable that ρ depends on r, and hence M/m should depend on tuner setting. That such is the case is shown by Fig. 2; also an experimental curve of $(M-m)/(M+m)$ *vs.* r showed a linear relationship.

[2] C. G. Montgomery *et al.*, *Microwave Measurement Techniques* (McGraw-Hill Book Company, Inc., New York, 1947), Chapter 11.

3. Phase

The discussion has been hitherto confined to the amplitude of the received wave. By the equipment in Fig. 4 one may measure the phase, the procedure being to observe the micrometer setting for minimum received power. A detailed description of the equipment and its use for dielectric measurement is given in reference 2. Proceeding as in (1) we find that the interaction of sheet and antenna should lead to a small phase error; and this error should show a nearly sinusoidal dependence on d with amplitude proportional to ρR and with period $\lambda/2$. Experimental verification is given in Fig. 5, where the amplitude of the experimental curve also leads to a value of ρ commensurate with those found in Section II.2.

4. Arbitrary Incidence

In practical work the dielectric sheet frequently makes an angle with the axis of the system. According to the present theory, interaction between antenna and sheet should be much reduced in these circumstances, since the reradiated wave is not reflected back into the r-f line (Fig. 6). That such is the case has been experimentally verified in that M/m and the phase error in Fig. 4 decrease with increasing angle. Closer investigation of the type summarized in Section III.4 shows that a normalized curve of $(M-m)/(M+m)$ *vs.* angle should approximate the secondary power pattern of the antenna, and this too has been experimentally verified.

In the general situation of Fig. 6 one finds that Eq. (3) is still approximately correct, provided T^2 is taken as the power transmission coefficient of the sheet at the given angle of incidence. Equipment based on these considerations can be designed to give rapid measurements of transmission *vs.* angle. A comparison of theory and experiment is given in Fig. 7 for a single sheet and for a more complicated configuration used in radome construction.

Arbitrary incidence introduces one possibility not found when the incidence is normal, *viz.*, the transmitted wave is elliptically polarized, in general, if the incident polarized is neither perpendicular nor parallel to the plane of incidence. This is true even for a homogeneous dielectric sheet, the mechanism being different from that for quarter-wave plates in optics. Insofar as it concerns the dielectric sheet the theory is not relevant to the present discussion;[3] what is of interest here is the effect on the receiving antenna. Assuming that both antennas are linearly polarized, one finds theoretically that a polar plot of $1/(P)^{\frac{1}{2}}$ *vs.* an angle specifying receiver polarization, will be an ellipse similar to that representing the locus of the E-vector in the elliptically-polarized wave produced by the sheet. Experimental verification is given in Fig. 8. Additional verification of this aspect of the theory is given in Fig. 9, where the experimental ratio of minor to major axis *vs.* angle of incidence is compared with the theoretical values as calculated from the direction of transmitter polarization and the electrical characteristics of the sheet.

[3] R. M. Redheffer, "Elliptical polarization produced by streamlined radomes," R. L. Report 483–13 (Feb. 12, 1946).

III. TRANSMITTING ANTENNA AND PLANE SHEET

1. Matched Line, Normal Incidence

For the over-all reflection into the line of a transmitting antenna and sheet arranged as in Fig. 1a the methods used to derive (1) lead to

$$\omega e^{i\omega'} = re^{ir'} + \frac{Rt\tau \exp[i(4\pi d/\lambda + t' + \tau' - \rho')]}{1 - R\rho e^{4\pi id/\lambda}} \qquad (4)$$

if the primed letters denote the phase shifts associated with the corresponding unprimed amplitudes. This gives

$$\omega = \frac{Rt\tau}{(1 - 2\rho R \cos 4\pi d/\lambda + \rho^2 R^2)^{\frac{1}{2}}} \qquad (5)$$

for the important case in which the line is matched, $r=0$. The equation is nearly the same as (1); it leads, for example, to (3) with M, m representing the extreme value of ω, and to

$$Rt\tau = 2Mm/(M+m) \qquad (6)$$

instead of (2). As in (1), the theory here too predicts nearly sinusoidal variation with period $\lambda/2$, and both properties have been extensively verified experimentally. The only exception occurs when a metal sheet is placed close to an antenna large compared to the wave-length. In such cases the variation is not always sinusoidal, though there is

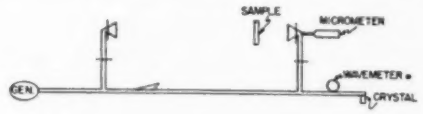


FIG. 4. Equipment for phase measurement in free space.

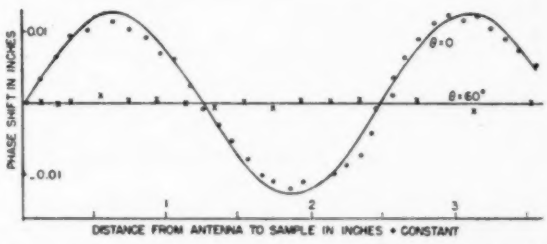


FIG. 5. Phase as measured by the equipment of Fig. 4 *vs.* position of sheet. Points are experimental at $\lambda=1.25$ cm curve is best-fitting theoretical.

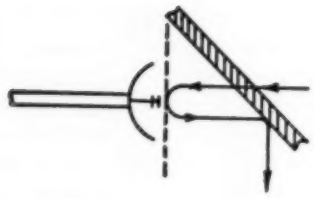


FIG. 6. Decreased interaction of sheet and antenna at non-normal incidence.

still $\lambda/2$ periodicity (Fig. 10). This anomalous behavior was investigated more fully; the curve shape duplicates itself accurately as d increases but depends critically on angle.

To measure reflection coefficients, compare a metal sheet or dielectric sheet of known reflection with the unknown in the same position. Since a complete curve is required to find the maximum and minimum reflections by the standing-wave technique, direct application of Eq. (6) is not convenient. Instead we take ω for d and for $d+\lambda/4$ with known and unknown, then use the ratio of the averages. From (4) the error committed is of the order of $(R\rho)^2$ and negligible in ordinary practice. In a series of 25 measurements of the same sample with five different antennas and various values of d, the maximum deviation from the theoretical reflection was 2.5 percent of the theoretical value; the average deviation was 1.3 percent. Graphical comparison of experiment and theory is given in Fig. 11.

If the unknown reflection is very small, so that serious error is introduced by a non-zero value of r, the proper procedure is to move the sheet with fixed probe position on the standing wave detector, taking maximum and minimum readings. This method, which lends itself particularly well to free-space procedures, cannot be pursued in detail here; it is fully described elsewhere, however.[4]

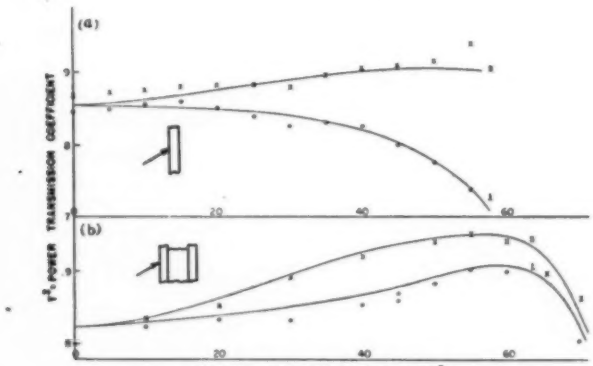


FIG. 7. Theoretical and measured transmission *vs.* angle at $\lambda=9.1$ cm and 10.0 cm. (a) Single ¼" plywood sheet, with polarization perpendicular and parallel to plane of incidence; (b) Same, sandwich construction with Hycar core and Fiberglas skins. Curves are theoretical from independent measurement of dielectric constant, loss, and dimensions.

These results are for amplitude. For phase, the minimum position follows the sheet along, as the latter is moved, the ratio of the two motions being proportional to the ratio of air and guide wavelengths. This conclusion, evident theoretically, has been experimentally verified for several antennas and wave-lengths; it is valid for arbitrary incidence and non-planar surfaces, as well as for plane sheets at normal incidence.

2. Fraunhofer Theory

We have assumed heretofore that the sheet is in the Fresnel region of the antenna, so that the transmission coefficients t, t' are substantially independent of the distance d from antenna to sheet. When this condition is not satisfied the theory of images suggests that the reflection will be inversely proportional to d, measured from some suitable point, if d is sufficiently large. Thus, the average reflection should be constant when the sheet is in the Fresnel region, and should decrease as $1/d$ when it is in the Fraunhofer region, a conclusion which is verified by the curves of Fig. 10, and by a number of similar curves obtained for various antennas. The envelopes were found from the equation

$$\text{envelope}=a/(b+d) \tag{7}$$

with the constants a, b adjusted for optimum fit. The value of b gives the effective center of the feed, that is, the point from which the rays appear to diverge; hence it must be the same for the upper and lower envelope.

A more detailed theory is obtained by combining inverse-distance attenuation with the results of the foregoing section. Thus, the apparent value of ρ for the Fraunhofer region, which is only a second-order effect in any case, may be assumed to decrease as $1/d$ with the same effective center as that found in the above operation on ω. This assumption leads to a modified form of (7) with constants a, b, c, and e,

$$\omega=\frac{a}{b+d}[1+c\sin(4\pi d/\lambda+e)], \tag{8}$$

if we note that, for the Fraunhofer case, R is always small. Equation (8) is verified in Fig. 12.

The effective center of the antenna has been found by several methods, and the results agree with each other and with the value of b for optimum fit in (7). One may also consider the critical distance at which the Fraunhofer theory begins to be valid. Generally speaking, the Fresnel theory as outlined above is most accurate for highly directive

[4] R. M. Redheffer and Yael Dowker, "The measurement of small reflections," R. L. Report 483-10 (February 6, 1945).

antennas, e.g., paraboloids, which give an approximately plane phase-front across the aperture. The larger the antenna compared to the wave-length, the greater is the distance d throughout which the Fresnel theory applies. Conversely, if the antenna is non-directive, e.g., very small in wave-lengths, then the Fraunhofer theory agrees with experiment even for small d, the required distance becoming larger with increased directivity or with increased ratio of antenna size to wave-length. This range of validity for the two theories is to be expected *a priori* from the usual behavior of Fresnel and Fraunhofer diffraction.

3. General Case, Normal Incidence

Proceeding now to the case in which the antenna is not matched to free space, we plot the complex number $\omega e^{i\omega'}$ on the so-called Smith chart, that is, the imaginary part *vs.* the real part as in the classical Argand diagram. It is known that one gets a circle and the details will be omitted. We ask, "Given the radius a_0 of the circle for a matched line ($r=0$), what is the radius a_r for the general case?" This formulation is appropriate because a_0 was investigated above. Neglecting the over-all reflection ρ of antenna-plus-tuner we get

$$a_r=a_0(1-r^2) \tag{9}$$

for the dependence of radius on mismatch. To obtain the dependence on c, the distance to the center, one makes a slightly different assumption, *viz.*, that the reflection ρ of antenna alone, without tuner, is zero. This gives

$$a_r=\frac{1+a_0^2}{2a_0}-\left(\left(\frac{1-a_0^2}{2a_0}\right)^2+c^2\right)^{\frac{1}{2}}. \tag{10}$$

Two geometrical interpretations are shown in Fig. 13, the first by Rieke, the second by Silver. As a consequence of the simultaneous validity of (9) and (10) the ratio of radii of the circles obtained for two different dielectric sheets will be independent of mismatch. This agrees with experiment.

The assumptions used to derive (9) and (10) were chosen differently in each case, the object being to minimize the effect of the reflection ρ, which is small but depends in an unknown manner on tuning. With these particular forms of the deviations it has been shown that omission of ρ probably has negligible effect on the final answers. We remark that to assume a lossless network for antenna-plus-tuner is incorrect theoretically and experimentally.

Verification of the above theory is given in Figs. 14 and 15. The mismatch in Fig. 14 was introduced by varying just one stub of a lossless tuner, whence it follows from simple theory (neglecting ρ) that

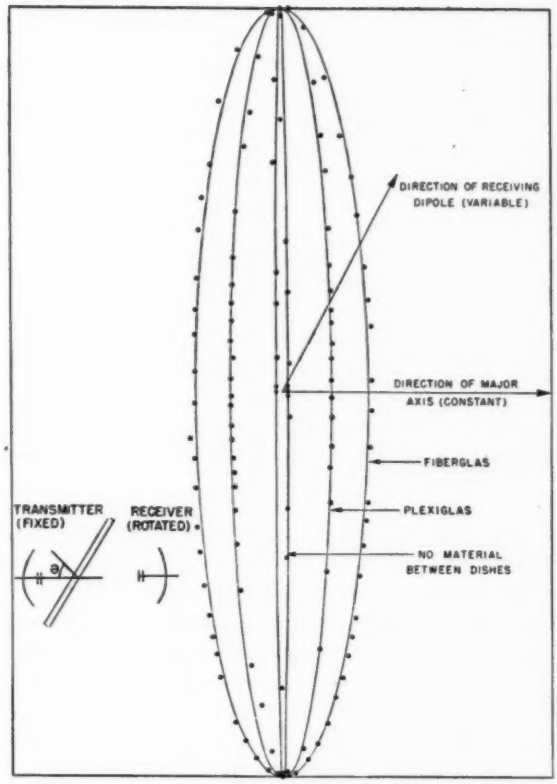


FIG. 8. For elliptical polarization of receiving wave, polar plot of 1/(received amplitude) as radius *vs.* direction of receiving dipole as angle. Points are normalized experimental, curves are best-fitting ellipses; $\lambda=9.1$ cm, $\theta=70°$.

the circle should behave precisely as shown in Fig. 13b. The initial mismatch $re^{ir'}$ moves along a circle, the center $ce^{ic'}$ along an ellipse. This is verified experimentally, and the curve obtained for a given mismatch is accurately a circle, as predicted. The equations for dependence of radius on initial mismatch or on distance to center are verified in Fig. 15. Each circle used for an experimental value

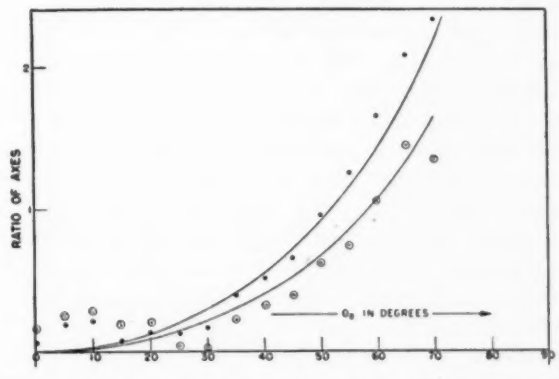


FIG. 9. Ratio of minor to major axis of ellipse *vs.* angle of incidence, when incident polarization makes an angle of 35° with plane of incidence. The points • are experimental for Fiberglas, ⊙ for plexiglas; curve is theoretical from independent measurement of dielectric constant and thickness. Equipment as in Fig. 8.

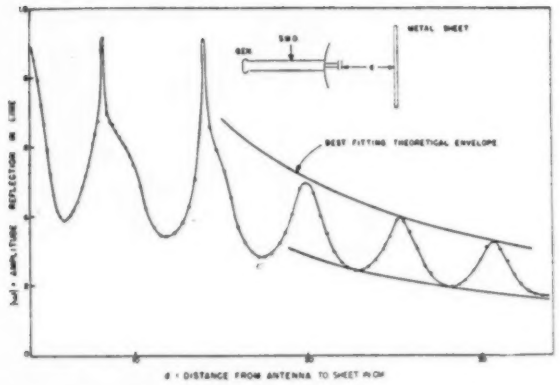


FIG. 10. Reflection *vs.* distance for metal sheet, showing transition from Fresnel to Fraunhofer region. Points are experimental for an 8″×2″ *f* paraboloid at $\lambda=9.1$ cm; envelope is best-fitting theoretical.

of a_r was determined by nine points on its circumference, as in Fig. 14, and many precautions were used in the experimental procedure.

The above assumes t, τ independent of d, that is, it assumes the case which has been designated as the Fresnel case. In the Fraunhofer case the radius of the circle will vary according to (7) or (8), so that a spiral is obtained. For moderately large antennas, however, the pitch is so small that it may often be neglected, when the variation of d gives only one or two rotations around the circle. In this case the only effect is to change a_0, and hence the theory of the present section is valid without modification.

4. Arbitrary Incidence

If the sheet is not perpendicular to the electrical axis of the antenna, the theory must be somewhat modified. Temporarily assuming a matched line and a negligible value for ρ, we note by the method of images that a plot of amplitude reflection ω *vs.* angle should be proportional to the secondary power pattern of the antenna, whenever the sheet is far enough away from the image to be in the Fraunhofer region (Fig. 16). If the antenna is directive, the same method may be used in the Fresnel region, the procedure being analogous to the use of Fraunhofer diffraction in the optical problem of the spectroscope although one is really working in the Fresnel region of both lens and grating. That amplitude reflection *vs.* angle should approximate the secondary antenna pattern has also been shown in other ways, which suggest that it is qualitatively correct even for non-directive antennas. The theory is confined to the maximum of the pattern and does not apply to the side lobes. Fortunately these have no importance for the present application; in the region of side lobes the reflection is generally too small to measure.

Before proceeding to experimental verification we must consider the effect of r and ρ, which have been neglected up to now but which may lead to large experimental errors none the less. A non-zero ρ produces sinusoidal variation similar to that noted above at normal incidence, except that the amplitude is reduced just as it was for the transmitting antenna. The effect may be canceled by averaging two measured reflections for which d differs by $\lambda/4$. In this way one obtains the result that would have been found with $\rho=0$. The dependence on r is likewise the same as for normal incidence; not only does one obtain a circle but the radius varies as predicted by Eqs. (9) and (10). In the Fraunhofer region too the effects are similar to those for normal incidence. It suffices, then, to consider only the case of a matched line, $r=0$, and to eliminate the effect of ρ by a quarter-wave displacement.

Turning now from theoretical to experimental results we find that amplitude reflection *vs.* angle indeed approximates the secondary power pattern near the peak for a wide variety of radiating systems, independently of the wave-length λ or of the distance d from antenna to sheet. Examples are given in Fig. 17, where each point is the average of two measurements with a quarter-wave change in d. As in Section III.1 anomalous effects are sometimes noted with a metal sheet. Although the curve accurately fits the secondary pattern near the peak, it may show a hump, similar to a side lobe, when the angle increases. This effect is noted only when the sheet is close to the antenna; with increasing distance the hump becomes a shelf, then gradually disappears, until finally there is agreement for all angles giving a reflection large enough to measure.

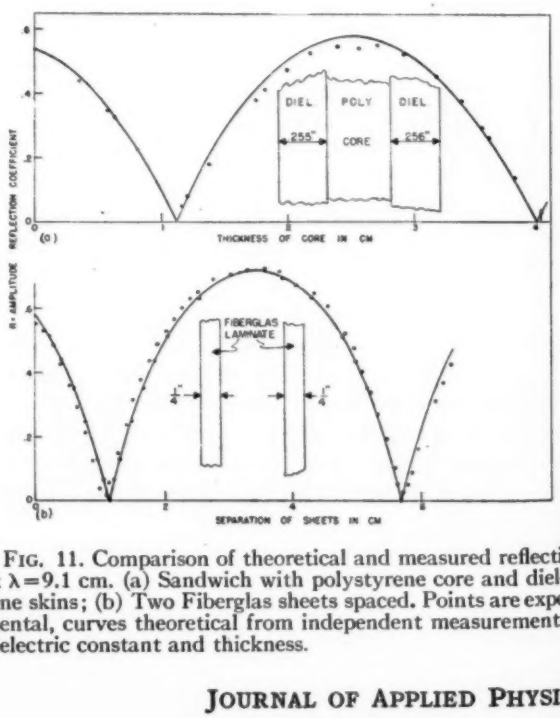


FIG. 11. Comparison of theoretical and measured reflection at $\lambda=9.1$ cm. (a) Sandwich with polystyrene core and dielectene skins; (b) Two Fiberglas sheets spaced. Points are experimental, curves theoretical from independent measurement of dielectric constant and thickness.

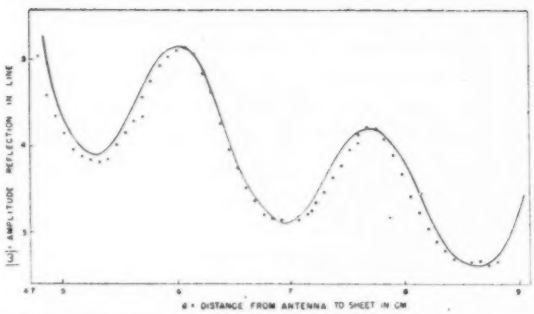


FIG. 12. Reflection *vs.* distance in Fraunhofer region, showing detailed comparison with theory. Points are experimental for a horn with $2\times2''$ aperture at $\lambda=3.2$ cm; curve is best-fitting theoretical.

It is assumed that the data are normalized, or, in other words, that the reflection into the line from a sheet of known properties has been given in advance. The ratio of the average reflection in the line to the reflection coefficient of the sheet is $t\tau$, which the present theory suggests will be about equal to the f-factor of the antenna, i.e., to the ratio of its gain to that of the same aperature uniformly illuminated. This is experimentally verified, but as to order of magnitude only; in typical antennas the value of $t\tau$ is between 0.65 and 0.75, while the f-factor is between 0.50 and 0.70. The theoretical value as determined from Eq. (11) is 0.70.

IV. TRANSMITTING ANTENNA AND CYLINDRICAL SURFACE

1. General

For a dielectric sheet in the form of a general cylindrical surface in front of a transmitting paraboloidal antenna, we require a theoretical expression giving the resulting reflection back into the line. This is a fundamental problem in radome design; it allows one to determine the so-called *pulling* of the magnetron. The effect of interaction between surface and antenna is to be canceled as in the case of the plane sheet, and the effect of fixed mismatch is likewise accounted for by the above theory. It suffices, therefore, to compute the average reflection into the line, with the antenna matched to free space.

As in Section III.4, we use Fraunhofer theory even though in the Fresnel region. The illumination across the surface is assumed equal to that across the antenna aperture, with the appropriate phase modification and with an amplitude change to take account of the reflection coefficient at the various angles of incidence. The radiation received by the antenna is computed by finding the Fraunhofer pattern of the surface thus illuminated and taking the amplitude in a direction parallel to the antenna axis. This procedure, which neglects all scattering except that in the direction of maximum gain, is valid only when the antenna pattern is sharp compared to that of the surface; in particular it is incorrect for a plane sheet at arbitrary incidence.

For a typical directive antenna of radius a the primary pattern is approximately

$$f(x)=[1-(x/a)^2]^2 \qquad (11)$$

as can be seen by starting from the known pattern of a double dipole feed and using geometrical optics. A similar equation[5] leads to antenna gains, side lobes, and half-widths comparable to those found in practice. While these quantities depend critically on $f(x)$, moreover, the reflection in the line is relatively insensitive. Hence a large percentage error in $f(x)$ is permissible if the shape near the peak is correct.

2. Circular Cylinder

For a circular cylinder of radius p concave toward the antenna and centered on the axis, the equation of the surface is

$$s(x)=(p^2-x^2)^{\frac{1}{2}}\cong\text{constant}-x^2/2p. \qquad (12)$$

We replace the dielectric reflection, which depends on angle, by a constant R—a procedure justified near the axis because of the near-normal incidence,

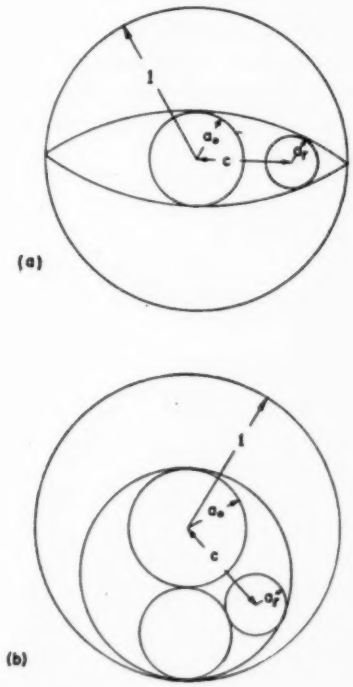


FIG. 13. Geometrical constructions for obtaining final radius a_r in terms of the initial radius a_0 and distance to center c.

[5] R. C. Spencer, "Paraboloid diffraction patterns from the standpoint of physical optics," R. L. Report T7, (Oct. 21, 1942).

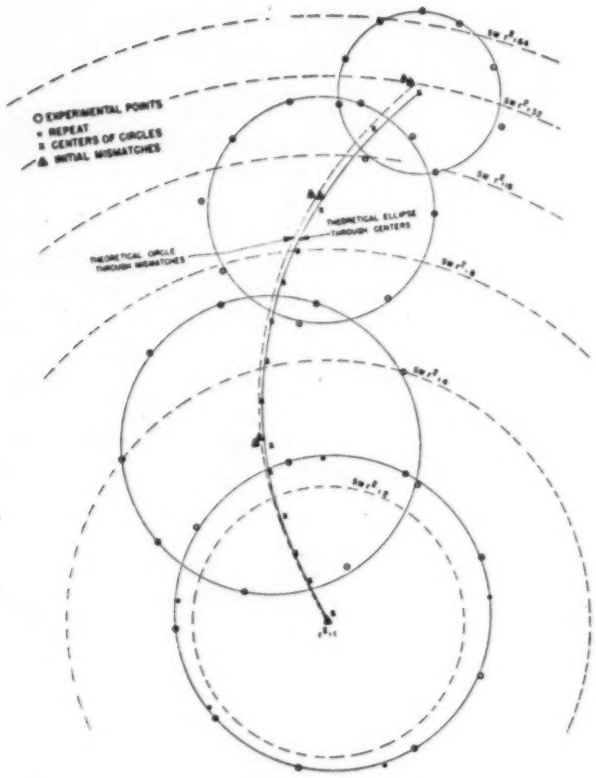


FIG. 14. Complex reflection as function of distance for various initial mismatches. Points are experimental with a 24″×7.2″ paraboloid, plexiglas sample; $\lambda=9.1$ cm.

and near the edge because of the small value of $f(x)$ in this region. Similarly the approximate member of (12) is permissible, the deviation being large only in areas of low illumination. Use of the above approach with (11) and (12) leads to an expression for ω which W. Ellis has expressed in terms of Bessel functions for $p=a$, and in terms of Fresnel integrals, with the approximation (12), for the general case. In Fig. 18 these are compared with the asymptotic form $\omega=21R(\lambda p)^{\frac{1}{2}}/32a\sqrt{2}\cong 6(\lambda p)^{\frac{1}{2}}/13a$ which Ellis has derived by letting $a/(\lambda p)^{\frac{1}{2}}\to\infty$, and with experiment. The points each represent about 45 reflection measurements; they were found by taking ω *vs.* d and integrating graphically for the average. The reflection was substantially independent of distance over the range considered ($15\text{ cm}\leq d\leq 60$ cm) and the curves had the same appearance as those for a plane sheet. This agrees with the theory. The data at $\lambda=9.1$ cm were obtained with a wide variety of cylinder radii, antenna sizes, and antenna shapes; hence nearly equal values of $a/(\lambda p)^{\frac{1}{2}}$ may represent very different experimental situations. The data are not assumed normalized. The theory takes us as far as possible, *viz.*, from the reflection coefficient of the sheet, computed as if it were plane, to the reflection into the r-f line. This reflection coefficient was independently determined; there is no "fitting" of experimental to theoretical values at any stage of the investigation.

The whole phenomena can be satisfactorily described in terms of the one variable $a/(\lambda p)^{\frac{1}{2}}$, a fact which, as L. J. Chu has pointed out, is explained by considering the first Fresnel zone on the cylinder. The more general theory shows the effect of neglecting zones other than the principal one, and it gives results for the transition region between cylinder and plane.

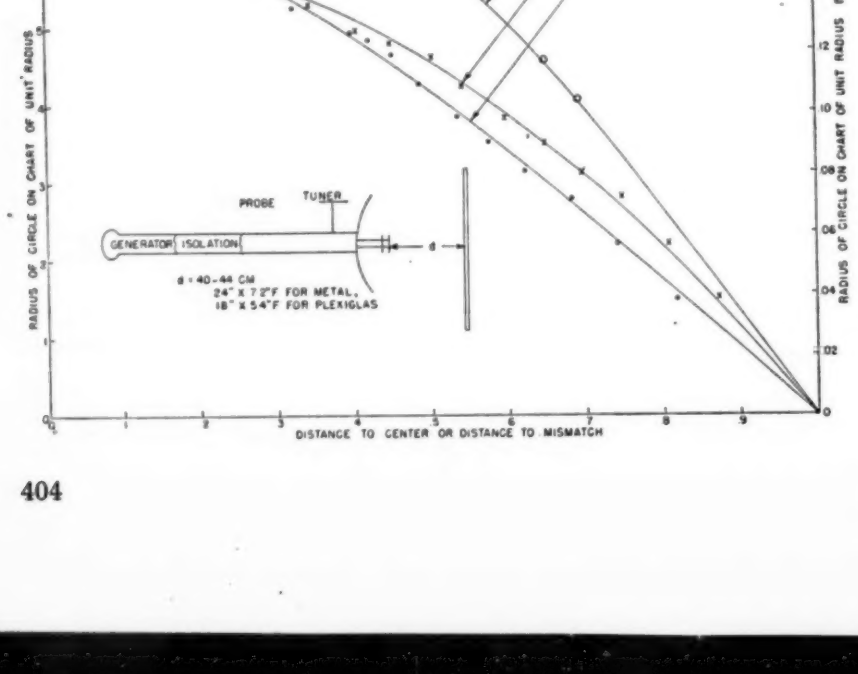

O, •, x EXPERIMENTAL
THEORETICAL
RADIUS VERSUS DISTANCE TO CENTER OR MISMATCH FOR PLEXIGLAS
RADIUS VERSUS DISTANCE TO MISMATCH FOR METAL
RADIUS VERSUS DISTANCE TO CENTER FOR METAL
RADIUS OF CIRCLE ON CHART OF UNIT RADIUS
RADIUS OF CIRCLE ON CHART OF UNIT RADIUS (PLEXIGLAS)
DISTANCE TO CENTER OR DISTANCE TO MISMATCH
PROBE
TUNER
GENERATOR
ISOLATION
d
d = 40-44 CM
24" X 7.2"F FOR METAL,
18" X 5.4"F FOR PLEXIGLAS

FIG. 15. Radius of circles like those of Fig. 14 *vs.* distance to center of mismatch for metal and dielectric sheets. Points are experimental, curves best fitting theoretical.

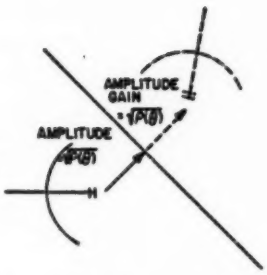


FIG. 16. Use of method of images to estimate effect of non-normal incidence.

3. Corrugated Surface

If the surface $s(x)$ is periodic with period L small compared to the antenna aperture, then we may replace $f(x)$ by a step function constant over each period. With this approximation ω *vs.* angle can be computed in terms of the pattern for a single period uniformly illuminated and a series of slits nonuniformly illuminated. It suffices to compute the envelope of the curve and the location of the maxima or minima; if these agree with experiment, then the detailed theoretical curve cannot fail to agree also. The angle of incidence θ_m for extreme reflection is given approximately by $\theta_m = \sin^{-1}(n\lambda/4L)(n=0, \pm 1, \pm 2, \cdots)$ where even n corresponds to a maximum, odd n to a minimum. The theory in this and the next section is based on these observations.

For the special surface of Fig. 19 experimental verification of this theory for normal and arbitrary incidence is given in Figs. 20 and 21. The data were taken as usual with ρ canceled by a quarter-wave motion of the surface, so that each point represents the average of two measurements of reflection. Altogether ten curves of the type shown in Fig. 20 were obtained for various strip widths and for polarization perpendicular as well as parallel to the corrugations. The agreement with theory was satisfactory except for the 1″ strips and perpendicular polarization, where edge effects are pronounced. Over 30 curves of the type shown in Fig. 21 were likewise obtained, for various values of d' and ϕ, and these too showed satisfactory agreement with theory. The agreement is better for the larger antenna, as one would expect from our initial assumption of high directivity. For the smaller antenna the maxima and minima of the curve are less well-defined; the irregularities are, so to say, smeared out. The effect is explained qualitatively by a more detailed theory which takes account of received waves not traveling parallel to the axis. In other respects this modified theory leads to substantially the same results as those found here, when the primary illumination $f(x)$ can be approximated near the peak by an error curve e^{-ax^2}; e.g., amplitude reflection *vs.* angle should still give the secondary power pattern.

Experiments of the type described here and in the next section show the extent to which Huygen's principle, in its simplest form, is applicable to systems with dimensions comparable to a wavelength. Generally speaking, fair accuracy is obtained down to about one wave-length, when the polarization is perpendicular to the air-dielectric edges, and to as low as a half wave-length when the polarization is parallel.

4. Strips

For a series of strips arranged as shown in Fig. 22a the above theory is verified in Figs. 23 and 24. At normal incidence the reflection is approximately proportional to the area occupied by the strips. In contrast to the behavior for metal, where the effect of conduction current appears to outweigh the effect of discontinuities at the edges, the reflection is higher for polarization perpendicular to the strips. The agreement with our scalar theory is also better for the parallel case. As usual, the curves of Figs. 23 and 24 are only representative of a much larger background of experimental data for various antennas and strip sizes. The other curves showed

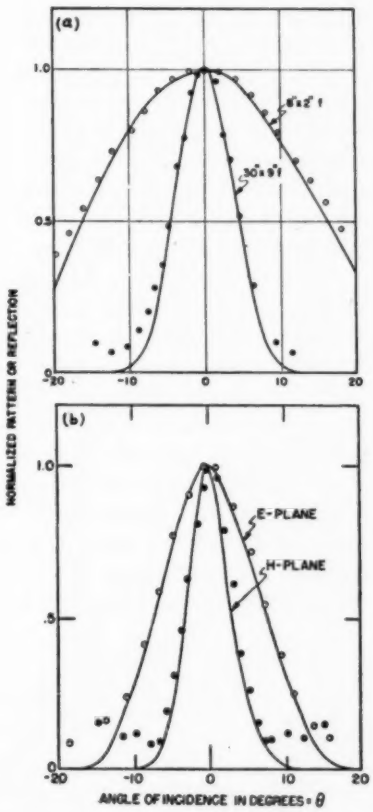


FIG. 17. Comparison of amplitude reflection *vs.* angle with secondary power pattern. (a) Result for two different antenna diameters, 8″ and 30″; (b) Result for E and H planes of a 35″ long by 12″ wide by 7.2″ f and cut dish. In each case the curve is secondary pattern and points are normalized amplitude reflection; $\lambda = 9.1$ cm.

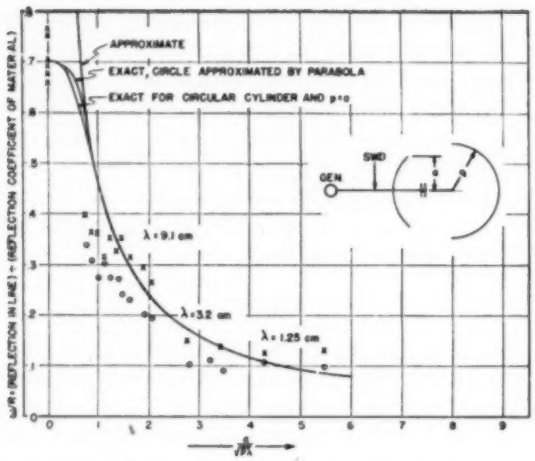


FIG. 18. Reflection from a circular cylinder into a paraboloidal antenna for various radii and wave-lengths. Points are experimental, curves theoretical.

much the same general behavior and equally good agreement with theory.

The foregoing equations do not depend on the antenna illumination $f(x)$. This simplification could be achieved because the period L was small compared to the antenna aperture. Turning now to the case in which this condition is not satisfied we find $\omega=(q/8)(15-10q^2+3q^4)\omega_0$ for the reflection from a strip of width $2aq$ centered on the axis and perpendicular to it (Fig. 22b). The result is ω_0 for $q=1$. Verification is given in Fig. 25, where we see again that the agreement is better, and the reflection lower, for polarization parallel to the strip.

If a sufficiently narrow strip be moved across the antenna aperture, as shown in Fig. 22c, its illumination and hence its reflection should be approximately proportional to the primary illumination $f(x)$. Upon re-entering the line the reflected wave experiences a further reduction $f(x)$, so that ω is approximately proportional to $f^2(x)$. Verification is given in Fig. 26. Examples for more general antennas were obtained[1] though no comparison was made with the primary illumination. The phase was likewise investigated; it was found to be nearly constant, as predicted by the theory, when the strip is in the Fresnel region. When in the Fraunhofer region, however, the antenna appears to be a point

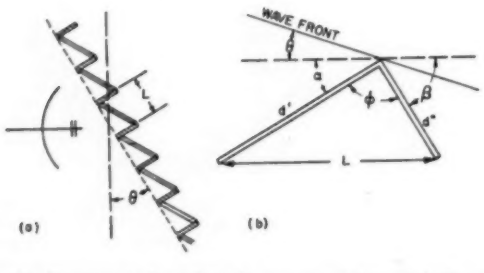


FIG. 19. Corrugated surface in cross section. (a) Complete surface. (b) Detail of one corrugation.

source; and hence the minimum position, as measured in the line, moves to account for the deviation between the rectilinear motion of the strip and the circular phase-front. This predicted behavior of the phase agrees with experiment, and leads to the same effective center as that found by the methods of Section III.2. Behavior of the amplitude in the Fresnel, transition, and Fraunhofer regions is shown in Fig. 27. The critical distance at which the beam starts to diverge agrees with the theoretical value (aperture area) f/λ given previously[6] and with the values at which the Fraunhofer theory begins to be valid in Section III.2. For a given distance from strip to axis in Fig. 27 the reflection is proportional to $1/d$ as in Section III.2; and the effective center for optimum fit again agrees with that found by the other methods.

5. Arbitrary Incidence

It has been assumed hitherto that the incidence is normal, in the sense that the cylinder element which intersects the axis does so at right angles. If the primary illumination $f(x,y)$ in the two-dimensional case may be written as $f(x)g(y)$, one finds that a plot of reflection *vs.* this angle should be approximately proportional to the same for a plane sheet, near the peak. Experimental verification is given in Figs. 28 and 29. Much additional work of this type has been done, and the result has been verified for other surfaces and materials (e.g., for a single strip and for a 12″ diameter sheet iron cylinder). The power standing-wave ratio (SWR)[2] in some curves decreased from a maximum of 1.25, and hence the experiment was somewhat exacting. Although measurement techniques have not been emphasized here it may be mentioned that similar difficulty is found for many of the experiments described, and hence that due care must be taken in the measurement procedure.

We note by the foregoing that reflection *vs.* angle for a plane sheet should approximate the secondary pattern; and for directive antennas the pattern in turn can be approximated near the peak by $\exp(4\theta^2 \ln 2/w)$, where w is the full width at half-power. This procedure leads to the results cited in the abstract. By such relations one may predict many aspects of radome performance without actually constructing a test radome; in conjunction with a so-called Rieke diagram, for example, one may predict the magnetron pulling as a function of the position in the scan. From this information and $i-f$ characteristics, is obtained the apparent pattern of the system with radome, that is, the signal strength on the scope *vs.* angle. Similarly one may investigate the apparent shift in cross-over

[6] S. J. Mason, "45° microwave reflector," Report 54-19 (Nov. 19, 1943).

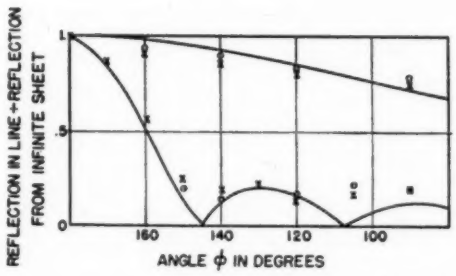


FIG. 20. Normal incidence, reflection *vs.* the corrugation-angle ϕ of Fig. 19. Points are experimental with polarization parallel to strips at $\lambda=9.1$ cm, × being for an $18''\times5.4''$ f antenna, ○ for a $30''\times9''$ f antenna. Curve is theoretical from independent measurement of dielectric constant and dimensions. Upper curve, $d=d'=1''$; lower, $=4''$.

point for lobe-switching systems, as a result of pulling on one lobe and absence of pulling on the other. The effect of overlaps or ribs in the radome construction can be estimated by the results in Sections IV.3,4. A full account of the procedure in such cases, with illustrative examples, is given in reference 1; since the emphasis here is on general results rather than specific applications the details are omitted.

V. PRESSURIZING

1. Design

Besides being useful in the study of radomes, these methods are helpful in the problem of antenna pressurizing, where one wishes to cancel the reflection as a result of the antenna with a second reflection caused by the pressurizing seal. This procedure was suggested by L. C. Van Atta. The advantages over the more obvious method of using an iris to match out the combined reflection of antenna and seal are threefold, *viz.*, the power-carrying capacity is not impaired; the fabrication is simpler; and the frequency sensitivity is reduced. It is therefore worth while to have some systematic method of designing a pressurizing seal in such a way as to lead to a matched line, without additional tuning devices.

The chief types of construction used in pressurizing are shown in Fig. 30, where we see that the sealing device may be a strip inside the guide or a cup surrounding the antenna. In the former case the procedure is very simple, as exact theoretical formulas are available for the reflection of a dielectric sheet in guide and for reflection *vs.* spacing of two objects. It suffices, therefore, to measure the initial mismatch r of the antenna alone, and to adjust the thickness of the dielectric for this same reflection. The position for minimum over-all reflection, which minimum will be practically zero, is then computed in an obvious way or found by trial.

The constructions of Fig. 30b and 30c offer

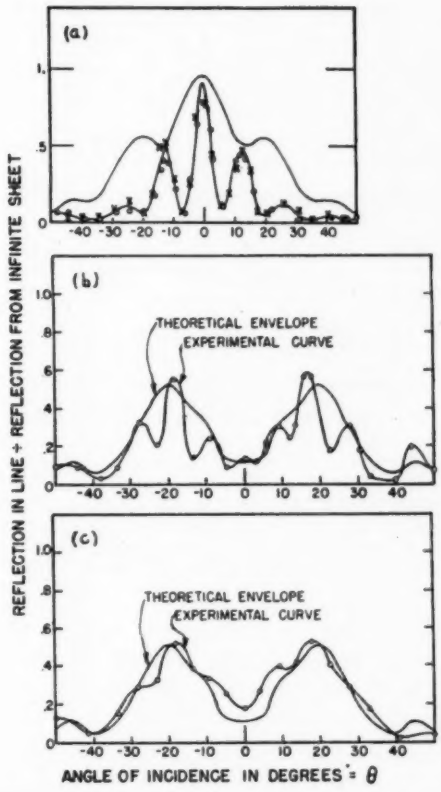


FIG. 21. Reflection *vs.* θ for a corrugated surface. (a) Reflection for $d'=d''=4''$, $\phi=160°$ in Fig. 19, and a $30''\times9''$ f paraboloidal antenna at $\lambda=9.1$ cm. Envelope is theoretical as in Fig. 20; vertical lines give theoretical location of maxima and minima. (b) Same, with $d'=d''=6''$ and $\phi=140°$. (c) Same as (b), except that antenna is $18''\times5.4''$ f.

advantages in power-carrying capacity and in frequency sensitivity, but do not permit a simple exact treatment of the sort just described. For these situations we introduce the concept of *effective reflection* into the line. If the initial mismatch of the antenna alone and the mismatch of

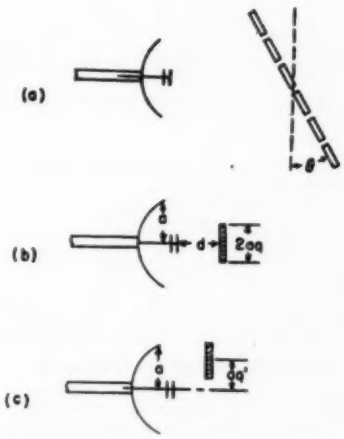


FIG. 22. Dielectric strips and antenna. (a) Series of strips at incidence θ; (b) Single strip centered on axis; (c) Single strip off axis. Figures show cross section only.

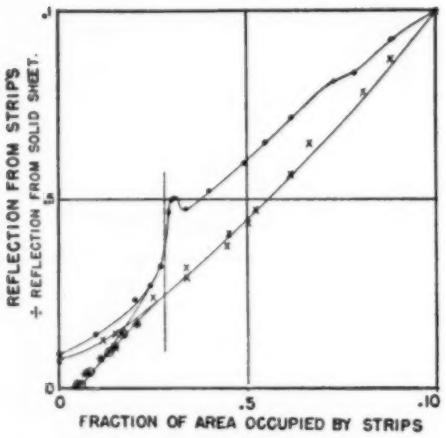


FIG. 23. Normal incidence reflection *vs.* spacing for a series of strips. The antenna is 30″×9″ *f* at λ=9.1 cm, and the strips 1″ wide. The points are for polarization perpendicular to the strips, × for parallel polarization. The cases of a space or a strip in the center are distinguished. The theoretical curve (not shown) is a 45° line through the origin.

antenna-plus-cup are plotted on the Smith chart, the effective reflection is defined as the vector joining these two points; and thus it represents that part of the observed mismatch which is due to the cup itself. To obtain a match it suffices that the effective reflection be equal in magnitude to that of the antenna alone and opposite in phase. The utility of this notion, which is also helpful in work with radomes, rests upon the following observations: For a given position of the cup the effective reflection is nearly proportional to its reflection coefficient R computed on the usual

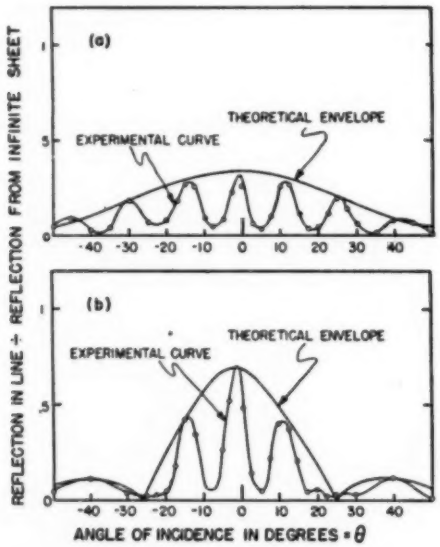

(a)
THEORETICAL ENVELOPE
EXPERIMENTAL CURVE
-40 -30 -20 -10 0 10 20 30 40
(b)
THEORETICAL ENVELOPE
EXPERIMENTAL CURVE
-40 -30 -20 -10 0 10 20 30 40
REFLECTION IN LINE ÷ REFLECTION FROM INFINITE SHEET
ANGLE OF INCIDENCE IN DEGREES = θ

FIG. 24. Reflection *vs.* angle from a series of strips. (a) Reflection for 2″ wide dielectric strips spaced 6″ apart, with a 30″×9″ *f* paraboloidal antenna and parallel polarization at λ=9.1 cm. Points are experimental, envelope theoretical; vertical lines give theoretical location of maxima and minima. (b) Same, with 4″ strips.

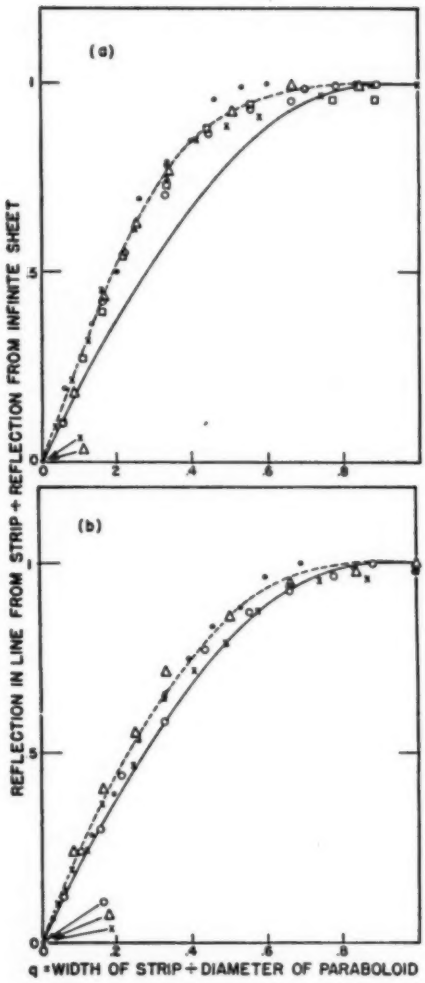


FIG. 25. Reflection of centered strip *vs.* width. The points Δ, ○, × and o are experimental for a 12″×3.6″ *f*, 18″×5.4″ *f*, 24″×7.2″ *f* and 30″×9″ *f* antenna, respectively, at $d=20$ cm; □ are for an 18″×5.4″ *f* antenna at $d=50$ cm. (a) is for perpendicular polarization, (b) for parallel polarization; dotted curve is empirical, solid curve theoretical. On all curves λ=9.1 cm.

plane-sheet basis and the phase of the effective reflection is nearly equal to a constant plus the electrical distance from antenna to cup. Both properties are consequences of the above theory, and have been extensively verified experimentally; examples for amplitude are shown in Fig. 31. The figure shows that effective reflection, for the strip inside the guide equals the theoretical reflection, in accordance with the above remarks. Also the reflection from the cylinder is larger than that from the plane sheet, as one would expect from the nearly cylindrical form of the phase-front near the antenna.

In view of these simple properties a systematic design procedure is to make a pressurizing cup similar in shape to the one desired but of a material such as plexiglas, that may be easily molded. The

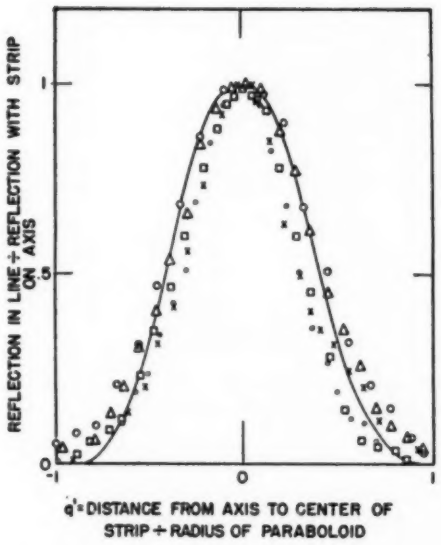


FIG. 26. Reflection from a narrow strip *vs.* its distance from the axis. Points as in Fig. 25, curve theoretical; 2.5″ strip, $\lambda = 9.1$ cm.

position is then adjusted for minimum reflection, whereupon the effective reflection is computed as the difference of the ω's before and after the cup was added. The reflection coefficient R of the trial cup is found from its thickness and dielectric constant, and finally the reflection coefficient necessary to achieve a match is computed by the proportion

$$\frac{\text{desired refl. coef. } R}{\text{refl. coef. of trial cup}} = \frac{\text{refl. from antenna alone}}{\text{eff. refl. of trial cup}}.$$

Once the desired reflection coefficient has been found, the thickness of the final pressurizing cup, which need not be of the same material as the trial cup, is found by the usual methods. The only change in the situation of Fig. 30c is that the phase now depends chiefly on the size of the cup rather than its position, and hence the new size as well as the thickness should be computed in advance. These methods have been repeatedly used with success.

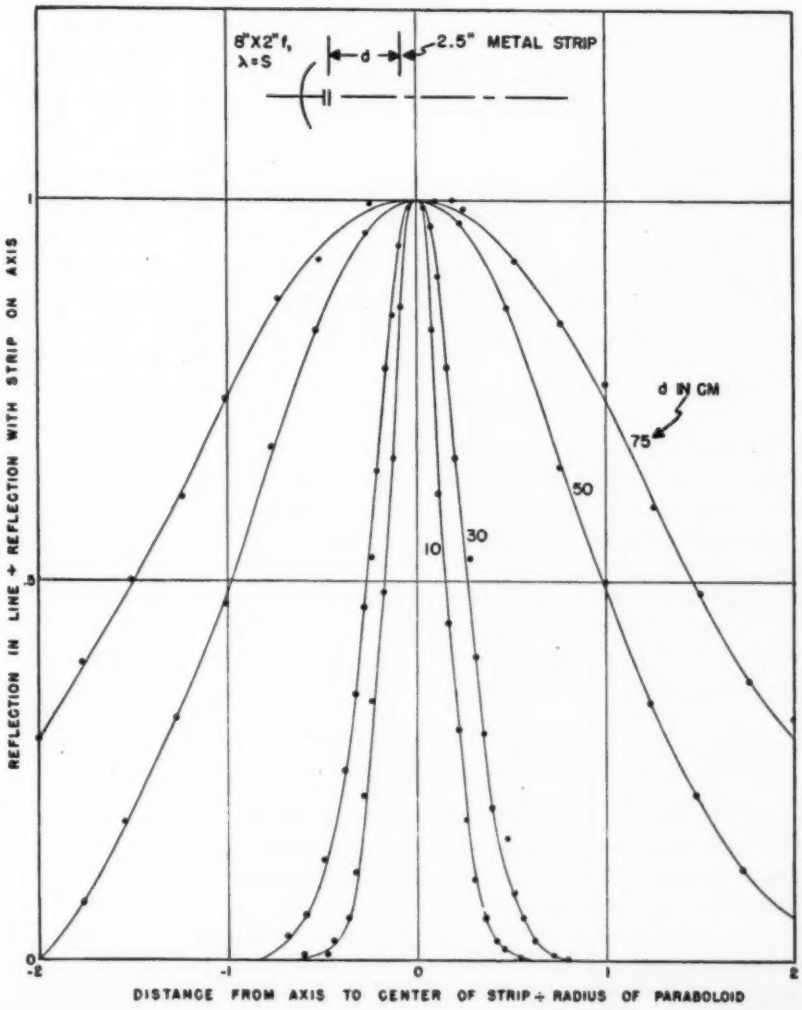


FIG. 27. Same as Fig. 26, but with an 8″×2″ f antenna, showing transition from Fresnel to Fraunhofer region.

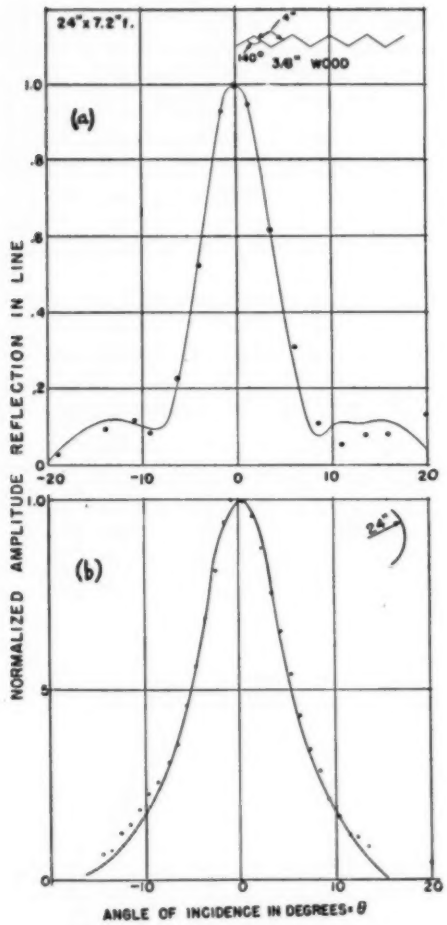


FIG. 28. Reflection *vs.* angle from cylindrical surfaces into a directive antenna compared with same for plane sheets. (a) Corrugated surface and 24″×7.2″ *f* antenna; (b) 24″ radius circular cylinder and 24″×7.2″ *f* antenna. In both cases curve is for plane sheet, points for cylinder; λ=9.1 cm.

2. Tolerances

Having found the optimum dimensions, we consider the effect of deviations due to manufacturing inaccuracies or variations of wave-length. For simplicity assume all amplitude reflections to be independent of cup position or wave-length, an approximation which is valid with slight error for the small variations encountered in questions of tolerance. With d as the difference between the optimum and actual position, we thus obtain

$$\begin{aligned} \omega &\cong [g^2 - 2gh\cos(4\pi d/\lambda) + h^2]^{\frac{1}{2}} \quad \text{(in general)} \\ \omega &\cong 2g\,|\sin(2\pi d/\lambda)| \quad \text{(optimum eff. refl.)} \end{aligned} \tag{13}$$

for reflection *vs.* position, size, or wave-length, as the case may be, in terms of the constants g, h depending on the antenna and cup under consideration. The guide wave-length should be used for λ in constructions like that of Fig. 30a. The tolerance in position depends on the actual value of the wave-length, while the frequency sensitivity depends on its derivative with respect to frequency; hence this construction shows higher frequency sensitivity but lower sensitivity to position than the others. Equation (13) has been experimentally verified with the frequency and cup thickness as the variables whose tolerances are being considered; examples showing the dependence on position are given in Fig. 32. The constants were determined for optimum fit, one constant being free for adjustment in each curve. For the optimum effective reflection, $g=h$, and in Fig. 30a, g equals the initial reflection r, while h equals the effective reflection of the strip. In Fig. 30b, g and h are proportional to but less than these values when dependence on position is being considered. This effect is due to the sides and ends of the cup, which suffer no relevant change in position as the cup is moved; in other words the part of the mismatch canceled by the sides and ends continues to be canceled after motion.

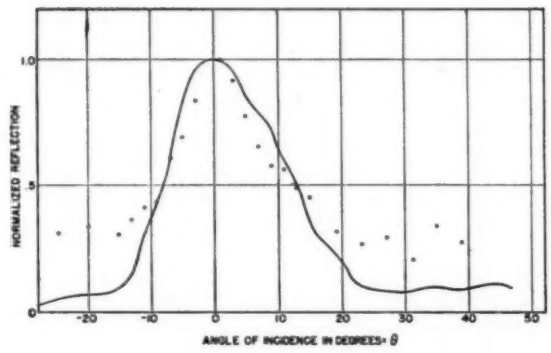


FIG. 29. Same as Fig. 28, but 15″ radius cylinder and 30″ csc² antenna, showing csc²-pattern plane.

We remark in passing that the pressurizing cup has effect on the antenna pattern, the change being not measurable for the situation of Fig. 30a and negligible in the other cases. On the other hand, the flange in Fig. 30b, which affects the match only slightly, may influence the pattern, the general effect being to reduce the width. Clearly, the optimum procedure is to adjust the flange for optimum pattern and then, with the flange soldered in place, to adjust the cup for optimum match.

A more detailed account of the material in this article, with much added experimental verification, is given in the Radiation Laboratory (R.L.) report cited as reference 1.

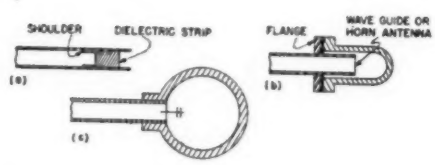


FIG. 30. Methods of pressurizing in standard use. (a) Dielectric strip or plug inside guide; (b) Cup and flange outside of guide or horn antenna; (c) Spherical cup for dipole antenna.

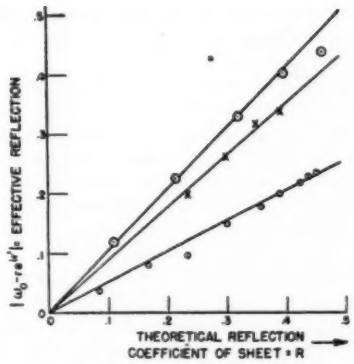


FIG. 31. Verification that effective reflection is proportional to theoretical plane-sheet reflection. The points ⊙ are for a strip inside the guide, × for a cylinder outside, and ○ for a plane sheet outside. Antenna is an open guide at $\lambda = 9.1$ cm.

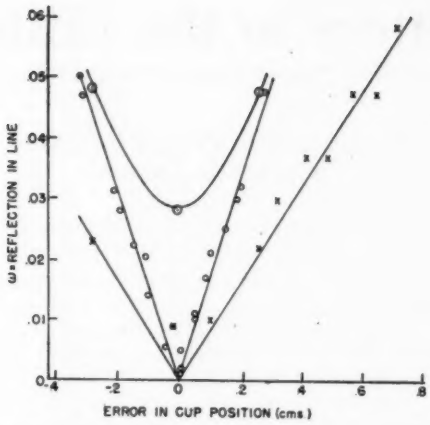


FIG. 32. Effect of errors in position of a pressurizing cup. The points × are for optimum cup thickness and a guide antenna; ○ are for optimum thickness and a 2.5″×8.5″ aperture horn antenna; ⊙ are for the horn with incorrect cup thickness. Curves are best-fitting theoretical.

ACKNOWLEDGMENT

Interaction of antennas and dielectric surfaces was apparently considered first,[7] where reflection at arbitrary incidence was measured by use of two antennas. A half-wave ripple was observed and ascribed to interaction of sample and antenna, as here. Since the work was concerned primarily with finding the reflection coefficient of the sample, no quantitative theory of this effect was included; but such work was undoubtedly carried out in the course of the research.

Most of our results for a receiving antenna were obtained in 1942, and those for the transmitter in 1943–4. The author was then at M.I.T. under direction of L. C. Van Atta and E. B. McMillan. The data of Fig. 32 were taken by S. J. Mason and C. G. Lincoln; the data of Fig. 5 are due Winkler. The theoretical curves in Fig. 7 were obtained by Y. Dowker. The data of Figs. 10 and 12 are due Steele, those of Figs. 23 and 24 to M. Furey. The author is grateful to B. Luz, M. Furey, and M. Hegarty for much of the work of computation, and to L. J. Chu for critical reading of this article.

[7] Pfister and Roth, Hoch:tech. u. Elekt:akus. **51**, 156 (1938).

Letters to the Editor

An Aberration Phenomenon in Electrostatic Lenses

SHIRLEY HARRISON
Physics Laboratories, Sylvania Electric Products Inc., Bayside, New York
January 7, 1949

THE gun structures employed in standard electrostatic cathode-ray tubes are designed in such a way that the stream of electrons emitted by the cathode can be focused to a small spot on a fluorescent screen when appropriate voltages are applied to the various electrodes in the system. Under these conditions, the electron lens element in the gun structure is relatively weak, but it is possible by changing the focusing voltage to increase the strength of the lens considerably. Operating an electrostatic gun at a fixed distance from a fluorescent screen, the change in focus is evidenced by the fact that such things as the shadows of various electrodes or the image of the cathode can be observed. The grid bias must be properly adjusted so that the contrast is good but the screen is not endangered.

With such a variation in the strength of the electron lens, say in a standard 7JP4 cathode-ray tube, the point of focus can be brought so far back into the gun structure that the shadows of the two pairs of deflection plates are seen within a brightly illuminated circular area, the outer edge of which is the shadow of the last round aperture. If under these conditions the lens is weakened, a dark area appears within the bright field between the shadows of the deflection plates. As the lens is made still weaker, this area grows in size, usually taking on a kidney shape. A photograph of this so called "kidney effect" in a 7JP4 tube is shown in Fig. 1. Further

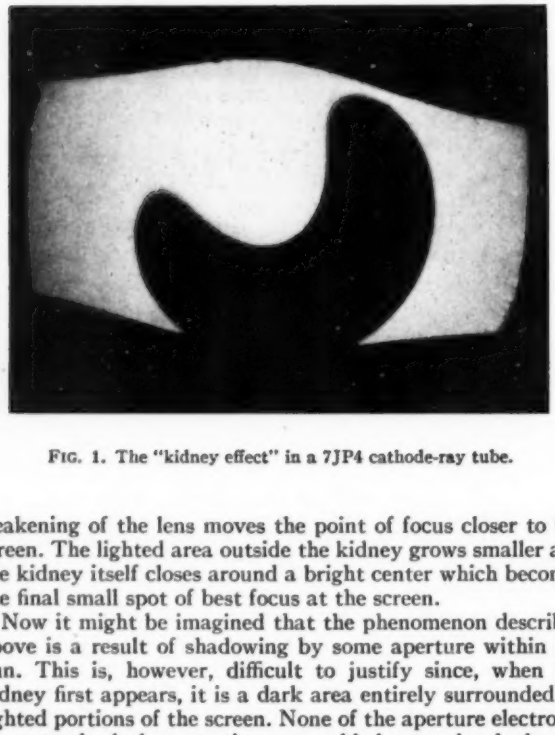


FIG. 1. The "kidney effect" in a 7JP4 cathode-ray tube.

weakening of the lens moves the point of focus closer to the screen. The lighted area outside the kidney grows smaller and the kidney itself closes around a bright center which becomes the final small spot of best focus at the screen.

Now it might be imagined that the phenomenon described above is a result of shadowing by some aperture within the gun. This is, however, difficult to justify since, when the kidney first appears, it is a dark area entirely surrounded by lighted portions of the screen. None of the aperture electrodes in a standard electrostatic gun could, by purely shadowing effects, intercept electrons somewhat near the axis and still permit those far off the axis to get through; yet only this kind of selective interception could produce the dark and light areas observed on the screen.

The appearance of the kidney within the shadow of the deflection plates furnishes some clue to its origin. At any earlier stage than shown in the photograph, there is an illuminated area outside the pair of deflection plates nearest the face of the tube. This is not true of the other pair of plates. It would seem then that the point of focus must, in this case, be located somewhere near the last round aperture of the gun, in the vicinity of the horizontal deflection plates. Further evidence to support this assumption was found during a study of current division in an experimental 7JP4 tube. With a fixed last anode potential and fixed grid bias, the focusing voltage was adjusted to vary the lens action from strong through weak and a minimum current to the last anode was observed just before the kidney appeared.

Assuming that the electrostatic lens in the gun structure is focusing the beam at a point close to the last round aperture, the focusing action is relatively strong, and, as is shown from lens studies in electron optics, a considerable amount of spherical aberration must be present. Admitting such a lens defect, an explanation of the kidney effect follows rather readily. Referring to Fig. 2, consider the paths of various

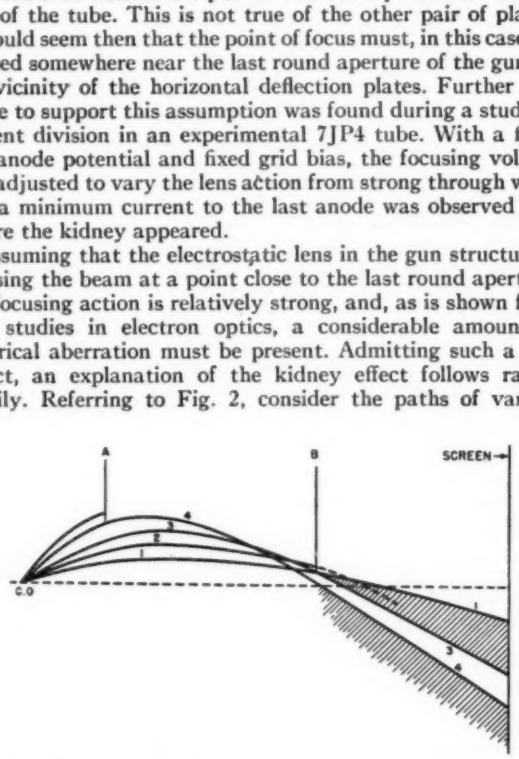


FIG. 2. The effect of placing an aperture near the point of focus of a lens with spherical aberration.

electrons leaving the cross-over, c.o., and passing through a focusing field in the region between A and B. With spherical aberration present, the point at which an electron path crosses the axis will depend on the angle at which it leaves c.o. Those electrons which travel in paths close to the axis will pass through the aperture at B and path 1 represents the limiting case. An electron leaving the cross-over with a slightly larger angle will follow such a path as 2 and will be intercepted by the aperture electrode at B. However, for still greater angles, there will be paths such as 3 for which the effect of the spherical aberration will be large enough to allow the electrons to pass through the aperture at B once more. Finally, the angle at which electrons leave the cross-over and still reach the screen is limited by an aperture at A, such as the stopping aperture in the first anode, and path 4 then represents one edge of the beam.

The paths 1, 3, and 4 are shown extended to the screen and to complete the picture a similar set of electron paths should be drawn below the axis, but these contribute nothing to the argument and would unnecessarily complicate the diagram. The dimensions are, of course, distorted for the purpose of explanation. Now from the discussion above, it follows that no electrons will reach the screen in the area shown shaded between the paths 1 and 3. Since the action of the lens is rotationally symmetric, this means that a dark ring should appear on the screen. However, it is reasonable to assume that some misalignment in the electrode structure accounts for the kidney shape of the area actually observed.

A light lens with large spherical aberration should produce an effect similar to that described above when used in such a way that an image of a point source is formed in the vicinity of a small round aperture. To confirm this, an experiment was performed to demonstrate the light optical analogy to the kidney effect observed in a cathode-ray tube. An ordinary reading glass was used to focus the light from a point source onto a plane containing a small round hole. A white screen to simulate the fluorescent screen in a cathode-ray tube was placed somewhat farther from the source than the aperture plane. If the system was adjusted so that a bright spot, concentric with and slightly larger than the round opening was observed at the aperture plane, a dark ring within a large bright spot appeared on the screen. Any slight shift of the aperture or misalignment of the lens caused the dark ring to become kidney shaped. Since the lens was not chromatically corrected, the dark areas were not black, but the shape of these areas was as predicted.

The main criterion for producing the dark ring in the light optical analogy seemed to be exact position of the aperture in the axis of the beam, while alignment of the lens with source was of secondary importance. Since in a cathode-ray tube it is extremely difficult to align the electrodes in such a way that the center line of the electron beam passes directly through the center of the last round aperture, the kidney-shaped area is usually observed.

Returning to the electron optical case, as the lens action grows weaker the spherical aberration decreases and thus the lighted area outside the kidney diminishes. For a very weak lens, only the paraxial rays pass through the aperture at B and electrons moving in paths further away from the axis are simply intercepted by the aperture electrode limiting the beam in such a way as to produce a well-defined spot.

The Absorption of Microwaves in Gases*

GEORGE H. MILLMAN AND RICHARD C. RAYMOND
Department of Physics, The Pennsylvania State College, State College, Pennsylvania
January 14, 1949

THE absorption by a number of gases of electromagnetic waves in the 3-cm wave-length region has been measured similar to the method reported by Hershberger.[1] The absorption cell was a rectangular wave guide, 10 feet in length. Microwaves were generated by a Raytheon type 723A/B reflex klystron and transmitted through the wave guide in the TE_{01} mode. A crystal rectifier and microammeter measured the transmitted power. The minimum measurable attenuation was 0.16×10^{-4} neper/cm.

Curves showing attenuation of NH_3 as a function of pressure were similar to those reported by Hershberger.[1,2] The absorption-pressure curves of the other gases mentioned in Table I followed this same general pattern.

The absorption coefficients of four of the gases investigated were compared at five different wave-lengths as given in Table I.

TABLE I. Absorption coefficients of gases in an unbounded medium at one atmosphere pressure and room temperature.*

Gas		3.230 cm	3.210 cm	3.201 cm	3.189 cm	3.180 cm
Acetaldehyde	—CH_3CHO	15.6	15.1	15.0	14.5	15.6
Ammonia	—NH_3	11.3	12.3	12.7	12.8	10.9
Ethyl Chloride	—CH_3CH_2Cl	7.5	7.8	6.7	7.9	8.7
Difluorochloromethane	—CHF_2Cl	5.3	6.1	6.0	7.3	7.3

* Units of absorption—10^{-4} neper/cm.

TABLE II. Absorption coefficient of methyl iodide and acetone in an unbounded medium at room temperature.*

Gas		Absorption	Wave-length	Pressure
Methyl Iodide	—CH_3I	1.4	3.227 cm	17.1 cm of Hg
Acetone	—CH_3COCH_3	1.8	3.231 cm	15.0 cm of Hg

* Units of absorption—10^{-4} neper/cm.

The results obtained are in close agreement with Hershberger's[2] except in the case of ammonia. At a wave-length of 3.18 cm, pressure of 1 atmosphere, he calculated an absorption coefficient of 7.2×10^{-4} neper/cm for CHF_2Cl and 8.7×10^{-4} neper/cm for CH_3CH_2Cl both of which correspond with the values given in the above table. However, for NH_3, he obtained an absorption coefficient of 15.5×10^{-4} neper/cm compared with our value of 10.9×10^{-4} neper/cm.

The maximum attenuation observed for methyl iodide and acetone is given in Table II.

Because of their low vapor pressures, a more detailed study of these gases could not be made with the present apparatus.

* This paper is from a thesis by George H. Millman submitted in partial fulfillment of the requirements for the degree of Master of Science at the Graduate School of The Pennsylvania State College.

[1] W. D. Hershberger, "The absorption of microwaves by gases," J. App. Phys. **17**, 495–500 (1946).
[2] J. E. Walter, and W. D. Hershberger, "Absorption of microwaves by gases II," J. App. Phys. **17**, 814–822 (1946).

Discussion of "Geometrical Factors Affecting X-Ray Spectrometer Maxima"

ROY C. SPENCER
Chief, Antenna Laboratory, Cambridge Field Station, AMC, USAF
Boston, Massachusetts
December 21, 1948

THE paper on geometry of x-ray spectrometers by Alexander[1] is of interest to the writer who has published a few brief notes on the same general subject,[2–7] i.e., the analysis and the correction or minimization of the effect of the apparatus on recorded curves.

As there appears to be an awakening interest in this field, let me summarize the several basic principles which govern such problems, including Alexander's special spectrometer problem:

First: The problem of vertical divergence is discussed in reference 2, pages 622–629, and in reference 5.

Second: The output function of a given apparatus (assuming that the superposition theorem holds) is the convolution (or Faltung) of the input function and the response of the apparatus to a unit impulse (an example of this is Fig. 4 of reference 1). As in the case of circuit analysis, we may separate the apparatus into uncoupled or loosely coupled stages, as, for example, the effects[6] of: (1) the vertical divergence of the slit system; (2) the amplifier system or its component stages; (3) the recording galvanometer; and (4) other operations of developing the film or analyzing its results by optical scanning or numerical averaging.

Third: Following the custom of the statisticians who describe the fundamental properties of their statistical distributions by listing their moments, we now proceed with the following moment analysis. The nth moment of the response $F(x)$ to a unit impulse is $\mu_n' = \int x^n F(x)dx$. The input function and the response of each stage to a unit impulse is normalized so that the zero moment or area for each is unity ($\mu_0' = 1$). It follows that the first moment μ_1' for the output function is the sum of the μ_1' for the input function and for each stage. This is equivalent to the theorem in circuit theory which states that the over-all time delay of the circuit is the sum of the individual time delays of the various stages. Of importance

in x-rays is the correction to the peak position due to the first moment of the vertical divergence.[2,5,6]

Fourth: On shifting the origin of each function to its own center of gravity (μ_1'/μ_0') we obtain a new set of "reduced moments," μ_n. These have the remarkable property that μ_2 for the output function is the sum of the individual values of μ_2. Thus the second moment or the square of the standard deviation is strictly an additive property. All equipment whose response curves are essential positive, such as spectrometer slits and critically damped galvanometers, have positive second moments. An underdamped galvanometer on an amplifier stage can have a zero or a negative second moment which can be used to balance out the positive second moment of the slit system.[7]

Fifth: The third moments are likewise additive. The third moment is a measure of the asymmetry of the curve. By reversal of the direction of recording, the asymmetry of the vertical divergence can be *either added to* or *substracted from* that of the amplifier and galvanometer system[6,7]

The writer has at times recommended that the reduced second and third moments of all components of a given apparatus be listed and then balanced out by mutual adjustment.[7] R. L. Murray[8] has applied this method to the first four moments of two coupled circuits. This procedure should be valid when the apparatus does not distort the input curve by more than a few percent, i.e., for low frequencies of recording. However, where there is large distortion or where the accuracy is limited by noise[3] (either circuit noise or the statistical errors due to a finite number of x-ray or light photons), then one should employ the methods used by Professor Norbert Wiener.[9]

[1] Leroy Alexander, J. App. Phys. **19**, 1068–1071 (1948).
[2] Roy C. Spencer, Phys. Rev. **38**, 618–629 (1931).
[3] Roy C. Spencer, Phys. Rev. **46**, 337(A) (1934).
[4] Roy C. Spencer, Phys. Rev. **48**, 473(A) (1935).
[5] Roy C. Spencer, Phys. Rev. **52**, 761 (1937). (Equation (4) should read $\mu_n = K^n \int_{-\infty}^{+\infty} \psi^n f(\psi) d\psi$.)
[6] Roy C. Spencer, Phys. Rev. **55**, 239(A) (1939).
[7] Roy C. Spencer, Phys. Rev. **60**, 172(A) (1941).
[8] Raymond L. Murray, "Operational analysis of coupled circuits," Master's thesis, University of Nebraska, Lincoln, Nebraska (1941).
[9] Norbert Weiner, *Expropolation Interpolation and Smoothing of Stationary Time Theories* to be published by John Wiley & Sons, Inc., New York, in April 1949.

Cathodic Vacuum Etching of Metals

DON M. MCCUTCHEON*
Process Engineering Laboratory, Ford Motor Company, Dearborn, Michigan
January 11, 1949

AT the 1944 Electron Microscope Society Meeting, in Chicago, during a discussion on etching techniques, the author made the suggestion that cathodic sputtering of metals may offer a means of preparing a fresh surface for microscopic observation.

We recently had an opportunity to investigate the sputtering or disintegration phenomena that occur at the cathode during a high voltage discharge in a reduced gaseous atmosphere. The original object of initiating a research project in this field was to investigate the possibility of treating contact points for an automotive voltage regulator. It was believed that the cathodic bombardment of contact-point surfaces would serve the dual function of removing any grease films, and also remove some of the highly flowed surface metal formed during manufacturing.

Our experience indicated that subsequent assembly operations vitiated any advantage of the cathodic treatment of contact points. Comparative tests between production cleaning methods and cathodic treatment showed only approximately 10 percent less failures for the cathodically bombarded samples. Further tests will be made.

The possibility of using cathodic bombardment as a means of producing etched structures in metals was investigated. An electrical contact point of silver-cadmium-nickel was investigated first, as the chemical-etching methods available were not satisfactory. Figure 1 shows the etch structure which was produced. Comparison of this pattern with the best chemically etched micrograph showed that the cathodic treatment brought out the true microstructure, but with greater detail and clarity.

Subsequent cathodic etching tests on steel forgings have shown that flow-lines can be brought out with a sharpness and clarity far exceeding anything that can be obtained by chemical etching. Flow-lines in steel samples have been produced by cathodic bombardment which could not be brought out by any metallographic method tried.

Pressure-rolling of bearings, particularly in the fillet area, is one method of introducing compressive surface stresses in order to increase the fatigue life. A Ford front-wheel spindle was experimentally pressure-rolled and examined by the usual microscopic techniques. No superficial layer of flowed metal could be detected by chemical etching techniques. Cathodic-bombardment etching, however, brought out the entire superficially rolled structure, showing individual flow-lines which indicated that deformation was only 0.005-inch deep, as shown in Fig. 2.

Some results have been obtained which indicate that grain structure can also be clearly shown by cathodic bombardment. An extensive program of investigation is under way to explore this phase of the method. Some comment has been made in the literature on the structures formed by cathodic disintegration, but the explanations offered are somewhat confused.[1]

The experimental technique for bringing out flow-lines in steel is as follows: Equipment used for metal evaporation techniques was adapted for sputtering by carefully insulating all metal surfaces and electrodes with glass-tubing or plates. The cathodic discharge should be confined entirely to the sample under examination. A cast aluminum disk, 3 inches in diameter and ¼-inch thick, is used as an anode. The cathode support, brought out through the top of the bell jar, is carefully insulated with glass-tubing. The sample is mounted in any convenient manner directly to the cathode rod, approximately 6 inches above the anode.

A diagram of the setup used for cathodic-bombardment etching is shown in Fig. 3. An argon atmosphere of approxi-

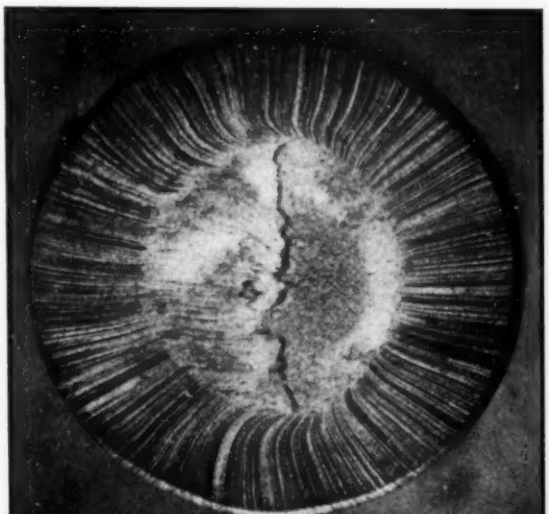


FIG. 1. Silver-copper-nickel contact point, etched by cathodic vacuum sputtering technique.

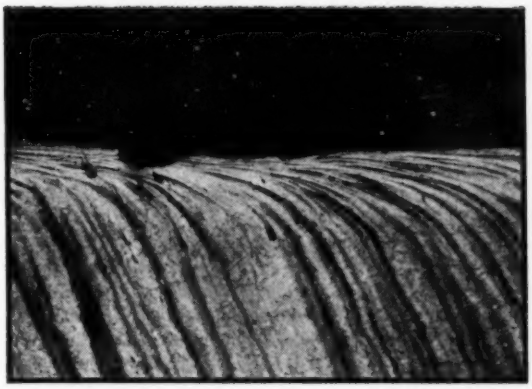


FIG. 2. Rolled fillet area of spindle, etched by cathodic vacuum sputtering technique.

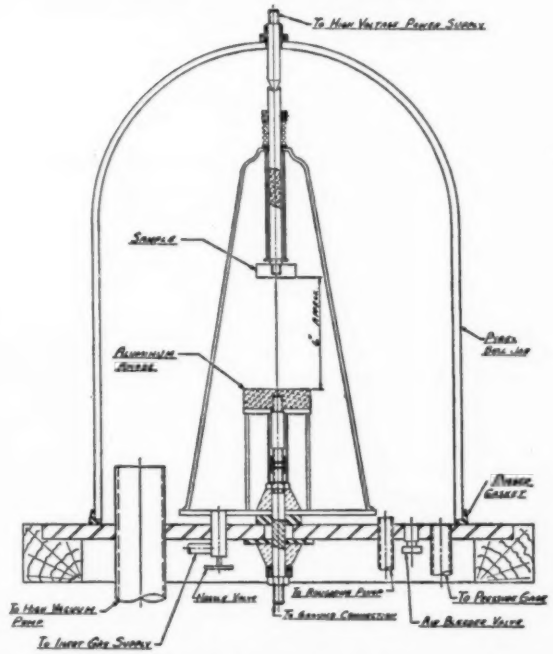


FIG. 3. Apparatus diagram.

mately 20 microns is maintained during the etching treatment. which takes from one to four hours to complete.[2] In the preliminary work, a 12,000-volt neon sign transformer was used with the secondary directly connected to the anode and cathode leads. The pressure in the bell jar was adjusted to give a current of approximately 20 milliamperes. With the current used during the etching process, it is estimated that the actual voltage between cathode and anode is only a few thousand volts, as the transformer used is current-limiting.

Mr. S. B. Miller produced the first sample etched by cathodic treatment in this laboratory, and Mr. William Pahl has carried out all subsequent experiments.

* Supervisor, Unit of Chemical and Metallurgical Research Head, Applied Physics Research.

[1] (a) A very extensive literature discusses the general subject of gaseous discharge, but the following publications show a few illustrations of etched patterns on metals produced by cathodic discharge; Beiträge zur Erklärung der Erscheinungen bei der Kathodenzerstäubung. Theodor Baum, Köln-Dellbrück. Zeits. f. Physik **40**, 686–707 (1927). (b) Dr. Cyril S. Smith, "Note on cathodic disintegration as a method of etching specimens for metallography," J. Inst. Met. **38**, 133–136 (1927). The present author's work differs from that reported in the references above in that flow-lines in forged metals can be brought out in addition to the usual microstructures of single or multiphased strain-free metals.

[2] Samples have been etched in less than 15 minutes by using a direct-current power supply.

Table of Interplanar Spacings for Different Target Materials for the Back Reflection Region

SUZANNE VAN DIJKE BEATTY
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
January 3, 1949

A TABLE has been computed to expedite the reduction of diffraction patterns obtained in the back reflection region and may be considered a continuation of Research Report R-94602-10-C.

The table gives d values for Mo, Cu, Co, Fe, and Cr targets for every tenth of a degree in 2θ, or five hundredth of θ, to four and five significant figures. The values of the interplanar spacings, d, are computed from the Bragg equation: $n\lambda = 2d \sin\theta$ where θ is the glancing angle, λ the wave-length of the diffracted target radiation, and n is taken as unity. The table is computed in Angstrom units for the $K\alpha_1$ radiation and factors are given for converting these spacings to those for $K\alpha_2$ and $K\alpha$. Conversion curves from $K\alpha_1$ to kx units are also given for all five target materials over the angular range.

The tables may be obtained directly from the author.

Pulse Emission Decay Phenomenon in Oxide-Coated Cathodes

G. R. FEASTER
Tube Department, Radio Corporation of America, Harrison, New Jersey
February 7, 1949

SEVERAL investigators have advanced hypotheses to explain the cathode fatigue effect which sometimes occurs during pulse-emission operation of barium-strontium oxide thermionic emitters. For example, Sproull[1] gives data on experimental diodes showing a decay in plate current of 20 to 40 percent in a time interval of 10 microseconds, and develops at length a theory based on the electrolytic migration and thermal diffusion of barium. E. A. Coomes[2] proposes an "electron reservoir" to account for the transient emission. In a private communication, Dow and Houghton of the RCA Rocky Point Laboratories have considered the possibility that the decay is a result of cooling caused by "electron evaporation." These hypotheses suffer from the common defect of being unable to account for tubes which do *not* show decay when tested under proper conditions.

In 1947, E. G. Widell and R. A. Hellar[3] found the short-time decay effect in very few tubes in a large series of tests in this laboratory. It was thought that the decay was due to some type of poisoning effect, but no further investigations

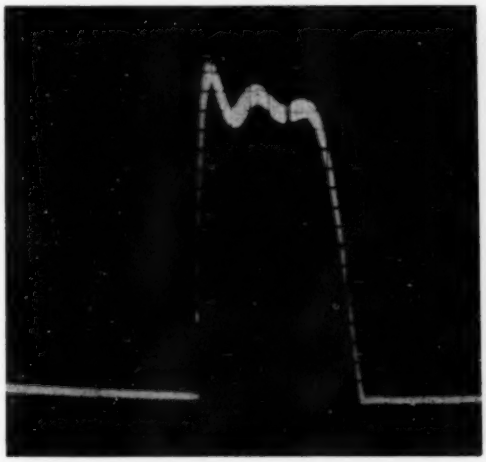


FIG. 1.

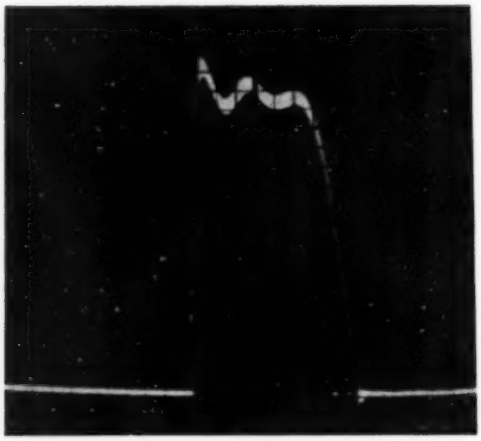


FIG. 2.

were made. More recent tests in this laboratory show that this short-time decay is an anomalous effect which is very infrequently found when testing receiving-type tubes with pulses up to 10 microseconds in duration. To ascribe this decay to a fundamental mechanism inherent in every barium-strontium oxide emitter would seem, therefore, to be a mistake. It should be noted that the experience of the M.I.T. Radiation Laboratory, as recorded by Glasoe and Lebacqz,[4] is that some tubes show a marked decrease, as much as 40 to 50 percent during a 2-microsecond pulse; others do not show the decay effect. They state that cathode fatigue is probably related in a complex way to the method of processing, inherent characteristics of the cathode material, and the manner in which the material is activated in the completed tube.

An investigation here,[3] using 10-microsecond pulses, has shown that decay in the temperature-limited emission is not a function of the ratio of barium oxide to strontium oxide. Other investigations show that the phenomenon does not correlate with the use of various types of cathode base metals. It has also been shown that damaging the emission of a cathode by exceeding the dissipation rating of the anode does not cause decay. However, a correlation with "poisoning" conditions has been established. Figure 1 shows a 10-microsecond pulse, at a repetition frequency of 100 per second, as observed on the synchroscope of the test set-up used when a pure resistance is substituted for a tube. The ripples result from the finite number of sections (5) in the pulse-forming network. Figure 2 shows the current pulse of a typical 5V4-G operating at a peak current density of 7.4 amp./cm²; no decay can be seen. Figure 3 typifies the result of adding 600 micrograms of sodium chloride to the anode of a 5V4-G before processing; this addition poisons the tube and causes it to show decay on pulsing. Identical results were obtained with a type 6AL5 tube operating at 6 amp./cm² approximately. The cathode temperatures in these tests were normal, i.e., about 1050°K. Sodium fluoride gives the same results as sodium chloride.

It has been shown by means of a tube with a rotatable anode that the decay effect is obtained only when the electrons bombard a portion of the anode containing a poisoning agent. No decay is noticed when electrons bombard uncontaminated portions. The question of whether oxygen, either free or in an oxide on the surface of the anode, will cause the decay phenomenon has not been settled and is now being investigated.

[1] R. L. Sproull, Phys. Rev. **67**, 166 (1945).
[2] E. A. Coomes, J. App. Phys. **17**, 647 (1946).
[3] E. G. Widell and R. A. Hellar, "The effect of composition of the coating on electron emission from oxide-coated cathodes," submitted for publication to J. App. Phys.
[4] Glasoe and Lebacqz, *Pulse Generators* (McGraw-Hill Book Company, Inc., 1948), p. 97.

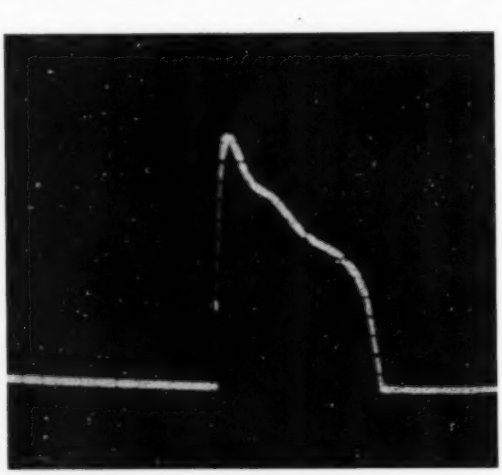


FIG. 3.

Running Striations and Multiple Electron Temperatures

N. L. OLESON* AND C. G. FOUND
Lamp Development Laboratory, General Electric Company, Nela Park, Cleveland, Ohio
February 7, 1949

IN the course of investigating the effect of the presence of argon on the electrical characteristics of the positive column of the low voltage d.c. mercury arc, we have found an hysteresis effect in the onset of running striations which has made possible probe measurements with and without striations, all other conditions being the same.

For relatively high argon pressures the striations are always present provided the arc current has been kept below a value i_{max}. They disappear, however, if the current is increased beyond i_{max}, and they will not reappear, even if the current is subsequently restored to its former value provided that it is greater than i_{min}. Only by decreasing the current below i_{min} can the striations be made to reappear and they will remain present until the current is again allowed to exceed i_{max}.

The experimental arrangement was as follows: A discharge tube, 3.6 centimeters in diameter and filled with 7 mm of spectroscopically pure argon plus mercury vapor, contained

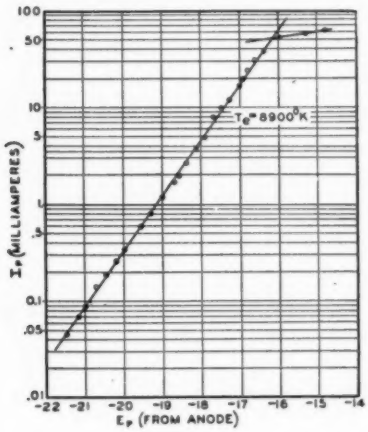


FIG. 1. Probe characteristic for mercury discharge containing 7 mm of argon at 0.4 amp., 74.5 volts, and 40°C. No oscillations.

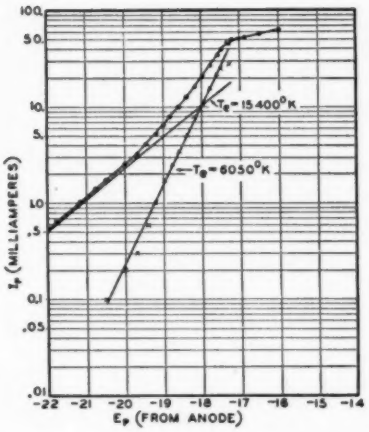


FIG. 2. Probe characteristic for mercury discharge containing 7 mm of argon at 0.4 amp., 75 volts, and 40°C. Oscillations 1000–1500 c.p.s. of 2.7 volts peak-to-peak amplitude.

two cylindrical tungsten probes each 5 mils in diameter and with an effective length of one centimeter, symmetrical with respect to the tube axis. They were approximately 30 centimeters apart and were located well within the positive column. Between the probes was a quartz window which was used to measure the intensity of the 2537A mercury line (using a caesium photocell). The cathode consisted of a well degassed oxide-coated filament of the type used in the General Electric sodium vapor lamp, and the anode was a nickel cylinder one inch in diameter and one inch long. The whole tube was immersed in a well stirred water bath, the temperature being kept at 40±0.5°C. The discharge was operated at each current for sufficient length of time to achieve stability as indicated by no change in the voltage across the tube. A cathode-ray oscilloscope was used to detect the oscillations, being connected between the anode and one of the probes. A slotted rotating disk was used to establish the fact that these oscillations were running striations.

It was found that if striations were present they would disappear if the discharge current was increased to 1.4 amperes. No striations were observed for currents greater than this value and they would not reappear until the current was decreased to approximately 0.25 ampere. Therefore it was possible to take probe data both with and without striations in the interval between 0.25 ampere and 1.4 amperes. No attempt was made to measure the frequency of the striations accurately, but it was between 1000 and 1500 cycles per second with a peak-to-peak amplitude of 2.7 volts for a discharge current of 0.4 ampere.

A Langmuir plot of the probe data obtained for a discharge current of 0.40 ampere and no striations is shown in Fig. 1. These show a Maxwellian distribution corresponding to an electron temperature of 8900°K over a range of 1000-fold. The measurements obtained with striations present and a current of 0.40 ampere do not follow a Maxwellian distribution. However, by successive approximations the results can be resolved into two groups having Maxwellian distributions as shown in Fig. 2.

The onset of striations had no effect on the tube voltage nor on the voltage gradient in the positive column as determined by applying a bucking voltage to the two probes to give zero probe current. The over-all intensity of the 2537A line remained unchanged. The average electron concentration at the tube axis as calculated from the break in the Langmuir plot is approximately the same in both cases, being 54×10^{10} electron per cc, if the electron temperature obtained from the experimental curve near space potential is used. The magnitude of the positive ion current to the probes remained unchanged, which would indicate that the positive ion current to the wall also did not change. The instantaneous values of these quantities would be expected, of course, to vary over the oscillation cycle.

It seems reasonable to associate the high energy electron group of electrons with the bright portion of the striation and the lower energy group with the dark region. The discharge has two modes of operation at a given current either of which can be produced at will.

* Present address: Naval Postgraduate School, Annapolis, Maryland.

Spherical Particles for Electron Microscopy

G. DAVID SCOTT
Department of Physics, University of Toronto, Toronto, Canada
February 7, 1949

AT the meeting of the Electron Microscope Society last September, Backus and Williams* pointed out the usefulness in electron microscopy of Polystyrene Latex, No. 580G, supplied by the Dow Chemical Company. This material is a water suspension of spherical particles of remarkable uniformity of size. The provisional value of the particle diameter was given as 2540±15A which was later increased to 2590 ±25A. Measurements taken recently using a sample of this latex from Lot 3584, sent to us in November through the courtesy of the Dow Chemical Company, indicate that though the particles on any mount are of nearly uniform size within the limits of error, the same average size is not obtained consistently.

The measurements were made by mounting the particles from a dilute suspension of the latex on Formvar replicas of a Baird grating of which the spacing was checked optically to be 1.695 microns. The actual spacing of each replica was determined by using a Hilger comparator to measure directly the distance between two identifiable groups of particles 100 to 200 grating spaces apart. Precise counting to a fraction of a space was done on a low magnification electron micrograph. This procedure gave a value, accurate to within ½ percent, of the average grating spacing of the replica over the region in question. It was evident from the results that shrinking of Formvar replicas often occurs, which must be allowed for in accurate work, though it is rarely greater than 3 percent. Higher magnification micrographs were then taken at a series of evenly spaced positions across the region selected, and the particle size determined relative to the replica spacing. The average value of the particle size found in this way does not depend on exact repetition of instrumental conditions and is independent of local variations in the replica.

The diameter of the particles was also determined by mounting them on a piece of very fine flat galvanometer suspension wire of which the width could be measured directly in the comparator, and then taking low and high magnification micrographs of particles on the edges of the wire. This method is not as satisfactory since it yields relatively few results in each case and there are instrumental factors involved. It did, however, confirm the sphericity of the particles.

The results showed that for any one mount, at least over a limited region, the average deviation of the particles from the mean size was little, if any, greater than the errors of measurement which were about 1 percent, but *from mount to mount differences in the mean size itself were quite appreciable*, the values ranging from 2500 to 3000A. This fact would seem to preclude as yet the possibility of giving a precise value to the average particle size. It was also observed that in many fields there appeared odd particles, occurring perhaps 1 in 10^4, smaller or much larger than the average size. (See Fig. 1(b) and (c).) Furthermore though most of the particles exhibited a high degree of sphericity, some were of irregular shape, some were fused with their neighbors, and some were contaminated

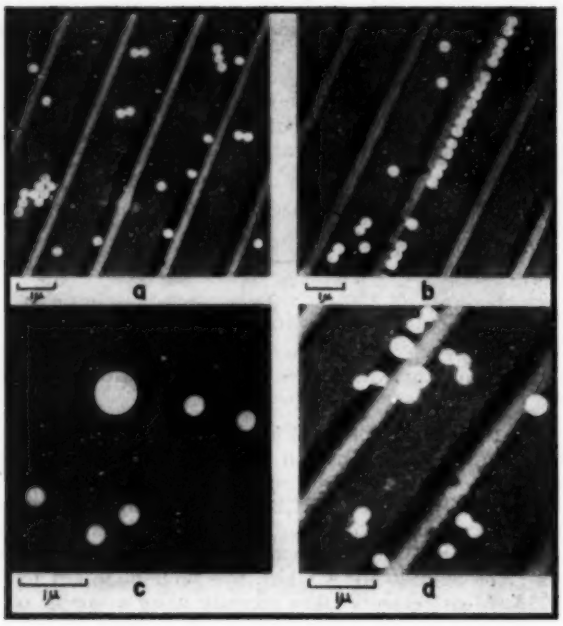


FIG. 1. (a) Particles mounted on Formvar replica of 1.7μ grating; (b) shadow cast mount, arrow indicates under-sized particle; (c) odd large particle with five normal particles on Formvar film (magnification same as for (d)) (d) particles of two sizes on grating replica.

or sunk into the Formvar substrate. As with many types of specimens, high intensity of the electron beam would often destroy the particles.

When a second drop of the diluted suspension was added to a previously prepared mount, two sizes of particles appeared. Usually the difference in size was just detectable with visual observation, but occasionally it was quite marked as in Fig. 1(d). Such observations suggest that the differences in mean particle size may depend on the condition under which the mount is made, e.g., the size of drop, its rate of drying, the way in which the drop wets the substrate etc., rather than on any inherent size distribution in the latex itself.

Because of their spherical shape these particles are indeed very useful in electron microscopy for obtaining shadow cast angles and film thickness, but it seems that some method of ensuring uniformity in their size is required before they can be used generally for precise size measurements. Standardizing the mounting technique might perhaps meet this requirement.

* R. C. Backus and R. C. Williams, J. App. Phys. 19, 1186 (1948).

Transonic Flow in a Shock Tube

D. K. WEIMER, C. H. FLETCHER, AND W. BLEAKNEY
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey

March 4, 1949

THE "shock tube" is a useful device for creating shock waves in the laboratory to study the properties of the waves themselves as well as other flow problems in the moving medium behind the shock front. It consists of a tube of uniform cross section divided by a thin diaphragm into two portions at different gas pressures. When the diaphragm is burst, a plane shock wave is propagated into the low pressure region setting the medium behind it in uniform motion. This kind of apparatus has been in use for a long time[1] but its potentialities for present day investigations are not fully appreciated. It was used a great deal during the war years at Princeton to study the properties of shock waves and the flow about obstacles,[2] especially pressure gauges. The device becomes especially valuable when suitable equipment is available for optical examination of the phenomena as has been shown in recent notes.[3]

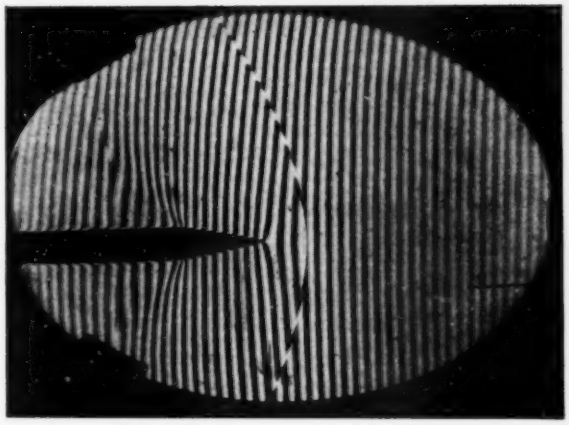


FIG. 1. Interferogram of two-dimensional flow of air at Mach number 1.0. The channel is about 4″×18″ and the wedge is ¼″ thick with a half-angle of 10°.

Figure 1 is an example of an interferogram taken with an instrument of the Mach-Zehnder type of the flow about a wedge for an incident flow of Mach number approximately 1. The density field may be obtained directly from the fringe shifts.[4] This picture was taken about 100 microseconds after the incident shock had passed the tip of the wedge so that this shock has disappeared from the field of view on the left. The flow is, of course, a function of time and such pictures have been taken showing the growth with time of the bow wave, the boundary layer and other features of the phenomenon at Mach numbers ranging from 0.9 to 1.4 where the bow wave becomes attached for this particular wedge. So far, only qualitative experiments have been made to show the possibilities of the method. The wedge was rather crudely made by hand. It is evident, however, that the incident flow is very uniform and the setup lends itself very well to quantitative measurements.

This work has been supported in part by the ONR contract N6ori-105 Task II.

[1] P. Vieille, Comtes Rendus **129**, 1228 (1899). Other early references are given by Dixon, Phil. Trans. **A200**, 315 (1903); W. Payman, Proc. Roy. Soc. **A120**, 90 (1928).

[2] G. T. Reynolds, OSRD No. 1519 (1943); Fletcher, Read, Stoner, and Weimer, OSRD No. 6321 (1946); W. Payman and W. C. F. Shepperd, Proc. Roy. Soc. **A186**, 293 (1946).

[3] W. Bleakney, Phys. Rev. **69**, 678(A) (1946); L. G. Smith, Phys. Rev. **69**, 678(A) (1946) and OSRD No. 6271 (1945). Mautz, Geiger, and Epstein, Phys. Rev. **74**, 1872 (1948); Fletcher, Weimer, and Bleakney, Bull. Am. Phys. Soc. **24**, 22, papers FA12 and FA13 (1949).

[4] Ladenburg, Winckler, and Van Voorhis, Phys. Rev. **73**, 1359 (1948).